Dhinaharan Nagamalai
Jan Zizka (Eds)

Computer Science & Information Technology

The Sixth International Conference on Computer Science, Engineering & Applications (ICCSEA 2016)
Dubai, UAE, September 24~25, 2016

AIRCC Publishing Corporation

Volume Editors

Dhinaharan Nagamalai,
Wireilla Net Solutions, Australia
E-mail: dhinthia@yahoo.com

Jan Zizka,
Mendel University in Brno, Czech Republic
E-mail: zizka.jan@gmail.com

ISSN: 2231 - 5403
ISBN: 978-1-921987-56-4
DOI : 10.5121/csit.2016.61101 - 10.5121/csit.2016.61106

Typesetting: Camera-ready by author, data conversion by NnN Net Solutions Private Ltd., Chennai, India

Preface

The Sixth International Conference on Computer Science, Engineering & Applications (ICCSEA 2016) was held in Dubai, UAE, during September 24~25, 2016. The Fifth International Conference on Signal, Image Processing and Pattern Recognition (SPPR 2016) and The Seventh International Conference on Ubiquitous Computing (Ubic 2016) were collocated with the ICCSEA-2016. The conferences attracted many local and international delegates, presenting a balanced mixture of intellect from the East and from the West.

The goal of this conference series is to bring together researchers and practitioners from academia and industry to focus on understanding computer science and information technology and to establish new collaborations in these areas. Authors are invited to contribute to the conference by submitting articles that illustrate research results, projects, survey work and industrial experiences describing significant advances in all areas of computer science and information technology.

The ICCSEA-2016, SPPR-2016, Ubic-2016 Committees rigorously invited submissions for many months from researchers, scientists, engineers, students and practitioners related to the relevant themes and tracks of the workshop. This effort guaranteed submissions from an unparalleled number of internationally recognized top-level researchers. All the submissions underwent a strenuous peer review process which comprised expert reviewers. These reviewers were selected from a talented pool of Technical Committee members and external reviewers on the basis of their expertise. The papers were then reviewed based on their contributions, technical content, originality and clarity. The entire process, which includes the submission, review and acceptance processes, was done electronically. All these efforts undertaken by the Organizing and Technical Committees led to an exciting, rich and a high quality technical conference program, which featured high-impact presentations for all attendees to enjoy, appreciate and expand their expertise in the latest developments in computer network and communications research.

In closing, ICCSEA-2016, SPPR-2016, Ubic-2016 brought together researchers, scientists, engineers, students and practitioners to exchange and share their experiences, new ideas and research results in all aspects of the main workshop themes and tracks, and to discuss the practical challenges encountered and the solutions adopted. The book is organized as a collection of papers from the ICCSEA-2016, SPPR-2016, Ubic-2016.

We would like to thank the General and Program Chairs, organization staff, the members of the Technical Program Committees and external reviewers for their excellent and tireless work. We sincerely wish that all attendees benefited scientifically from the conference and wish them every success in their research. It is the humble wish of the conference organizers that the professional dialogue among the researchers, scientists, engineers, students and educators continues beyond the event and that the friendships and collaborations forged will linger and prosper for many years to come.

Dhinaharan Nagamalai
Jan Zizka

Organization

General Chair

Natarajan Meghanathan	Jackson State University, USA
Dhinaharan Nagamalai	Wireilla Net Solutions, Australia

Program Committee Members

Abbas Akkasi,	Islamic Azad University of Bandar, Iran
Aditya Rockstar,	Jawaharlal Nehru Technological University, India
Adnan Albar,	King Abdulaziz University, Saudi Arabia
Ahmed arara,	College of Technology at Alkharj, Saudi Arabia
Ajith Nongmaithem,	Manipur University, India
Akhil Garg,	Nanyang Technological University (NTU), Singapore
Ali Hussein Mohammed,	Alexandria University, Egypt
Alireza Afshari,	Islamic Azad University, Iran
Amel Boufrioua,	University Constantine 1, Algeria
Amir baharvandi,	IEEE Transaction on Power Systems, Iran
Amit Chauhan,	Gujarat University, India
Amit Choudhary,	Maharaja Surajmal Institute, India
Anamika Ahirwar,	Rajiv Gandhi Technical University, India
Ankit Chaudhary,	Truman State University, USA
Barbaros Preveze,	Çankaya University, Turkey
Chin-Chih Chang,	Chung Hua University, Taiwan
D N Chandrappa,	SJB Institute of Technology, India
Dac-Nhuong Le,	Haiphong University, Vietnam
Damodar Reddy,	National Institute of Technology Goa, India
Daniel D. Dasig,	Jr., Jose Rizal University, Philippines
Dayakar C.V,	Anna University, India
Dinesh Chandra Jain,	S.I.R.T.S, India
Diptoneel Kayal,	West Bengal University of Technology, India
Doreswamy,	Mangalore University, India
Ehsan Saradar Torshizi,	Urmia University, Iran
Farhad Soleimanian,	Hacettepe University, Turkey
Gammoudi Mohamed Mohsen,	University of Manouba, Tunisia
Govindraj Chittapur,	Basaveswar Engineering College, India
Guo Yue,	Ningbo University of Technology, China
Hacene Belhadef,	University of Constantine 2, Algeria
Hamdi hassen,	Miracl laboratory, Tunisia
Hossein Jadidoleslamy,	MUT University, Iran
Irving V Paputungan,	Universitas Islam, Indonesia
Isa Maleki,	Islamic Azad University, Iran
Jaesoo Yoo,	Chungbuk National University, Korea
Jaesoo Yoo,	Chungbuk National University, Korea
Jamshid Aghaei,	Shiraz University of Technology, Iran

Technically Sponsored by

Networks & Communications Community (NCC)

Computer Science & Information Technology Community (CSITC)

Digital Signal & Image Processing Community (DSIPC)

Organized By

Academy & Industry Research Collaboration Center (AIRCC)

TABLE OF CONTENTS

The Sixth International Conference on Computer Science, Engineering & Applications (ICCSEA 2016)

The Fifth International Conference on Signal, Image Processing and Pattern Recognition (SPPR 2016)

The Seventh International Conference on Ubiquitous Computing (Ubic 2016)

EVALUATION AND STUDY OF SOFTWARE DEGRADATION IN THE EVOLUTION OF SIX VERSIONS OF STABLE AND MATURED OPEN SOURCE SOFTWARE FRAMEWORK

Sayyed Garba Maisikeli

College of Computer and Information Sciences
Al-Imam Muhammad Ibn Saud Islamic University
Riyadh, Kingdom of Saudi Arabia
maisikel@ccis.imamu.edu.sa

ABSTRACT

When a software system evolves, new requirements may be added, existing functionalities modified, or some structural change introduced. During such evolution, disorder may be introduced, complexity increased or unintended consequences introduced, producing ripple-effect across the system. JHotDraw (JHD), a well-tested and widely used open source Java-based graphics framework developed with the best software engineering practice was selected as a test suite. Six versions were profiled and data collected dynamically, from which two metrics were derived namely entropy and software maturity index. These metrics were used to investigate degradation as the software transitions from one version to another. This study observed that entropy tends to decrease as the software evolves. It was also found that a software product attains its lowest decrease in entropy at the turning point where its highest maturity index is attained, implying a possible correlation between the point of lowest decrease in entropy and software maturity index.

KEYWORDS

Software Evolution, Software maintainability and degradation, Change ripple-effect, Change Impact, Change Propagation

1. INTRODUCTION

After a software system is developed, there is a high possibility that it may undergo some evolution due to change in business dynamics, response to environmental change, bug fix exercise, improving design, preventive maintenance or intentional modifications for overall improvement of the performance of the software system. A small change in an object-oriented software system however, may produce major local and nonlocal ripple effects across the software system. The goal of software evolution is to explore and study ripple-effects and cumulative effects of changes over time; observing whether quality, stability and extendability of

Dhinaharan Nagamalai et al. (Eds) : ICCSEA, SPPR, UBIC - 2016
pp. 01–13, 2016. © CS & IT-CSCP 2016 DOI : 10.5121/csit.2016.61101

the software are affected as the software system evolves from one version to another. Considering the size and complexity of the modern software systems, tracking and discovering parts of the software impacted, risks associated with change, and consequences of a change cannot be overemphasized.

When used properly, change impact assessment can help in managing and assessing software maintenance risks, thereby providing guidelines for effective software evolution implementation. This project investigated six versions of JHotDraw, a widely used open-source Java graphics framework as a test suite. The novel idea about this project is that while similar research efforts used static data collection methods, this project applied dynamic data collection methods on the test suite software under study.

According to [1], software maintenance includes corrective, adaptive and perfective maintenance enhancements which are technically not a part of software maintenance but, being a post-release activity.

Identifying potential consequences of a change or estimating what needs to be modified to accomplish a change may be a daunting task. According to [2] when a software system undergoes modifications, enhancements and continuous change, the complexity of software system eventually increases, with a possibility that some level of disorder may be introduced, making the software system becoming disorganized as it grows, thereby losing its original design structure. Considering the size and complexity of the modern software systems, tracking the effect of the change, understanding change impact and what parts of the software are affected and possible risks associated with a proposed change and potential consequences (side-effects) of a change cannot be overemphasized. When used properly and effectively, software change impact assessment can proactively provide a means of managing software maintenance risks and help guide the implementation of the software change.

On the issue of measuring software degradation, [3 and 4] suggest the use of entropy as an effective measure, and opined that software declines in quality, maintainability, and understandability as it goes through its lifetime. This paper sets out to study six consecutive versions of JHotDraw, a matured and well-structured open source graphics software framework that has been widely used in many research projects as test subject software. Each of the test versions was subjected to dynamic profiling and tracing routine that collected data from which Shannon entropy and software maturity index were derived.

The goal was to observe the entropy level change, and whether there is any correlation between entropy and software maturity index as the software system evolves from one version to another.

2. RELEVANCE

According to [5], the two most common meanings of software maintenance include defect repairs and enhancements or adding new features to existing software applications. Another view expressed by [5] also opined that the word “maintenance” is surprisingly ambiguous in a software context and that in normal usage it can span some twenty-one forms of modification to existing applications.

According to [6], almost 50% of software life cycle cost is attributed to maintenance; and yet, relatively very little is known about the software maintenance process and the factors that influence its cost. Considering the cost magnitude associated with maintenance and the ever-increasing size and sophistication of modern day software systems; it is then clear that software maintenance cost decisions and associated evolution risks cannot be taken lightly.

3. RELATED STUDIES

In a software evolution research, [7], analyzed change of software complexity and size during software evolution process, and discussed the characteristics related to the Lehman's Second Law (Lehman et al., 1997), which deals with complexity in the evolution of large software systems and suggests the need for reducing complexity that increases, as new features are added to the system during maintenance activities. Also, [7] opined that addition of features leads to the change of basic software characteristics (such as complexity/entropy) in the system. Their paper used this change as a means to determine different stages of evolution of a software system, proposing a software evolution visualization method called Evolution curve (or E-curve).

Discussing software maintenance consequences, [5] also observed that in every industry, maintenance tends to require more personnel than those building new products. For the software industry, the number of personnel required to perform maintenance is unusually large and may top 75% of all technical software workers. The main reasons for the high maintenance efforts in the software industry are the intrinsic difficulties of working with aging software, and the growing impact of mass updates. In an empirical study conducted by [8], thirteen versions of JHotDraw and 16 versions of Rhino released over the period of ten years were studied, where Object-Oriented metrics were measured and analyzed. The observed changes and the applicability of Lehman's Laws of Software Evolution on Object Oriented software systems were tested and compared.

In a research paper, [9] presented how graph-based characterization can be used to capture software system evolution and facilitate development that helps estimate bug severity, prioritize refactoring efforts, and predict defect-prone releases. Also, [10] presented a set of approaches to address some problems in high-confidence software evolution. In particular, a history-based matching approach was presented to identify a set of transformation rules between different APIs to support framework evolution, and a transformation language to support automatic transformation.

In another paper, [11] compared software evolution to other kinds of evolutions from science and social sciences, and examined the forces that shape change, and discussed the changing nature of software in general as it relates to evolution, and proposed open challenges and future directions for software evolution research. From software evolution point of view, [12] described how and when the software evolution laws, and the software evolution field, evolved, and also addressed the current state of affairs about the validity of the laws, how they are perceived by the research community and the developments and challenges that are likely to occur in the coming years.

In contrast, this paper focuses on measuring software degradation in the evolution of six versions of a large-scale open-source software system with a special focus on investigating the introduction of disorder and observing the software maturity level as the software system evolves from one version to another.

4. METHODOLOGY

In order to explore and investigate the effect of change and its impact on the amount of disorder introduced as a software system evolves from one version to another, this study considers six versions of JHotDraw (JHD) as a test suite. These six versions were produced in a period of about five years (2006 to 2011), reflecting its natural evolution as new requirements were added, existing functionalities modified or enhanced, and some were deleted.

4.1 Test Program (JHotDraw)

JHotDraw is a very popular, mature and well documented widely used open-source Java-based graphics framework that has been used extensively in many software engineering research projects as a test suite. This framework provides a skeleton for developing highly structured drawing editors and production of document-oriented applications. The framework is known to be heavily loaded with numerous design patterns, developed based on the solid object-oriented principles, and based on the best software engineering practices.

To justify using the six different versions of JHotDraw in this research, we referred to some authors who have used them previously; this includes [7] and [8] where they recommended the use of JHotDraw as an Aspect Mining validation benchmark. Also, [13] and [14] used JHotDraw as a benchmark test suite in their research work. In addition, [8] used JHD as one of the test suites in his project.

Since JHotDraw is a mature and widely used test software programs, this research project also adopted it as a test program. It should be noted that, although there are ten documented versions of JHotDraw, seven versions are considered in this research study because the difference between earlier versions (7.0.6 and 7.07) is minimal as explained by [7]. To help us understand the chronological nature of the test program and its various versions, some characteristics details are presented in table 1 below:

Table I. Characteristics of the six versions of JHotDraw

Versions	Release Date	Size (MB)	LOC	No. Classes	NOM	No. of Attributes
Version 7.0.9	6/21/2007	11.2	52,913	487	4,234	1090
Version 7.1	3/8/2008	27.6	53,753	485	2,800	1087
Version 7.2	5/9/2008	22.6	71,675	621	5,486	1479
Version 7.3.1	10/18/2009	22.7	73,361	638	5,627	1516
Version 7.4.1	1/16/2010	22.6	72,933	639	5,582	1455
Version 7.5.1	1/8/2010	23.3	79,275	669	5,845	1599
Version 7.6	6/1/2011	23.5	80,169	672	5,885	1606

Seven different versions of JHotDraw are evaluated and tested (see table 1). Each of the versions of JHotDraw) were dynamically profiled and traced through the use of AspectJ run-timed weaver. (AspectJ runtime weaver is discussed in section 4.2). In order to maximize code coverage, forty-six of the major functionalities of each of the JHD applet versions were exercised as they execute. The granularity level adopted in targeting the various test program artifacts for data collection in this project is at the method level, rather than at class level.

One of the reasons for the choice is that methods in Object Oriented programming represent a modular unit by which programmers attribute well-defined abstraction of ideas and concepts. [15], defined methods in object-oriented paradigm as self-contained units where distinct tasks are defined, and where implementation details reside, making software reusability possible. According to [16], methods are less complex than classes, are easier to compare, and provide significant coverage and easy distinction, and have a high probability of informal reuse. [17] Observed that all known dynamic Aspect Mining techniques are structural and behavioral and work at method granularity level.

Event traces were dynamically collected as the test software versions were executed, with the AspectJ runtime weaver seamlessly running in the background. The runtime weaver has the capability to dynamically insert probes at selected points in the target test software (in this case class methods) at specify points known as (joinpoints), where all method executions were traced and data collected. In this project, we are interested in the sequence and frequency of calls, rather than method fan-in and fan-out. Frequency counts for each method calls were tallied, from which probabilities of method invocation were calculated and assigned.

Note that, since methods with the same name in different classes may be counted as one and the same, we left the class prefix along with method names to make sure that such methods are counted distinctly and correctly. Note also that duplicate method calls were left intact in the data collected, since removing such duplicate calls will distort the frequency counts of the method invocations.

The assigned probabilities represent the probability that such code units will be invoked as the system is run. It is from this frequency count that the entropy is calculated as the software changes from one version to another. The other metric used was software maturity index (SMI); this was derived from the static data collected from documentations produced by [15]. Explanation on how these two metrics are used are discussed in the next few pages.

4.2 Dynamic Data Collection tool (AspectJ Weaver)

AspectJ runtime weaver allows probes to be inserted at specific points of interest statically or dynamically when the software source code to be profiled executes. Code that allows observing tracing or changing the software source code is weaved according to the required action specified in what is called (pointcut). The weaved/inserted code logs the behaviors of the test software, track its actions based on the given behavior specified by pointcut; in our case, tracing and profiling each of the methods in our test software system as they are executed or invoked. AspectJ runtime weaver can be used to seamlessly and dynamically collect data on the test software as it executes.

The weaver evaluates the pointcut expressions and determines the (joinpoints) where code from the aspects is added. This may happen dynamically at runtime or statically at compile time. The runtime weaver then creates a combined source by weaving the source code of the aspects into the sources of the program under investigation. The generated program code is then compiled with the compiler of the component language, which is Java in our case.

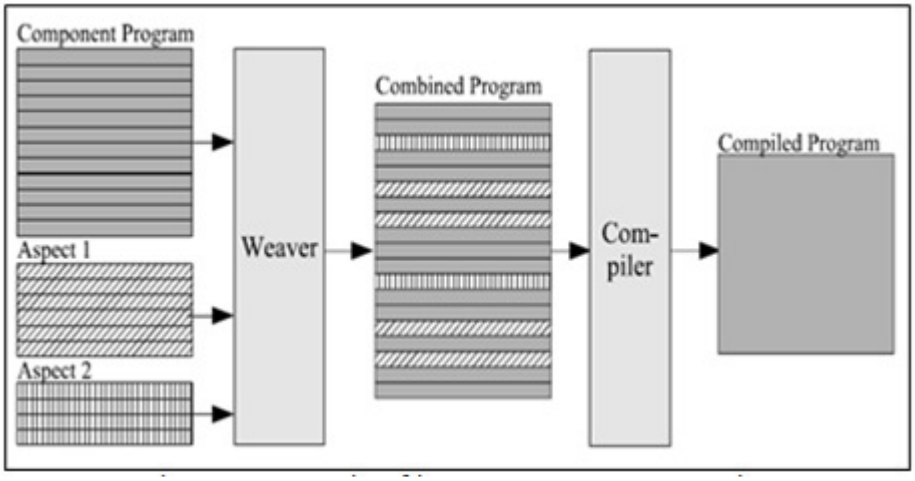

Figure 1. Example of how AspectJ Weaver works

4.3 Metrics derived from collected data

To assess, evaluate and study the nature of the test software as it evolves from one version to another, two software metrics were considered in this research project. Included are the Shannon's Entropy and Software maturity Index (SMI). These metrics were derived from the datum collected as the test programs run.

4.3.1 Shannon's Entropy

Within the context of software evolution, entropy can be thought of as the tendency for a software system that undergoes continuous change eventually become more complex and disorganized as it grows over time, thereby becoming more difficult and costly to maintain.

One of the metrics derived in this project is Entropy, with this metric; we will be able to find a way to assess whether the test software versions get degraded as they evolve from one version to another. According to [4], when investigating and studying the effect of a change in a software system, Shannon's equation may be better than complexity averaging. According to [1], in addition to measuring disorder introduced into software evolution, entropy also provides a measure of the complexity of the software system. [3], [20] stated that entropy can anecdotally be defined to mean that software declines in quality, maintainability, and understandability through its lifetime. For effective measurement and assessment of software degradation, [4] recommended the use of entropy for the study of software degradation.

Many variations of Shannon's entropy formula is presented in academic papers, but the generalized Shannon's entropy formula is expected as follows :

$$H_1 = -\sum_i p_i \ln p_i ,$$

Where

H = System Complexity Entropy,

p_i = Probability that method m_i in test software is invoked
i = Integer value *1, 2…j*, representing each of the categories considered

Note that the negative sign in the equation is introduced to cancel the negative sign induced by taking the log of a number less than 1.

As explained earlier in the introduction section, the entropy probability in this project is derived based on the method invocation frequency counts collected when the different versions of the test programs are executed and exercised. As an example of how entropy is derived in this project, consider the example of a software system **S** with three classes $\mathbf{C}_1$, $\mathbf{C}_2$, $\mathbf{C}_3$. Methods ($\mathbf{m}_{11}$, $\mathbf{m}_{21}$) are contained in $\mathbf{C}_1$, methods ($\mathbf{m}_{12}$, $\mathbf{m}_{22}$, $\mathbf{m}_{32}$) contained in $\mathbf{C}_2$, and ($\mathbf{m}_{13}$, $\mathbf{m}_{23}$, $\mathbf{m}_{33}$, $\mathbf{m}_{43}$, $\mathbf{m}_{53}$) contained in $\mathbf{C}_3$. The numbers shown beside class methods are representations of the frequency of method invocations when the test software was exercised.

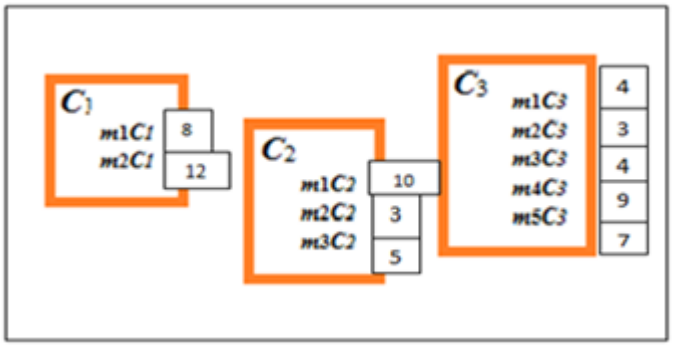

Figure 2. Example of method invocation from three different classes in (software **S**)

Based on the given example of the three classes and the associated method invocations shown in figure 2 above, we can construct probability required for the calculation of the entropies for all methods in the software being tested as shown in table 2 below

Table II. Example of calculation of probability of method invocation.

Classes	Invoked Methods	Invocation Frequency	Invocation Probability
C_1	m_1C_1	8	0.1231
	m_2C_1	12	0.1846
C_2	m_1C_2	10	0.1538
	m_2C_2	3	0.0462
	m_3C_2	5	0.0765
C_3	m_1C_3	4	0.0615
	m_2C_3	3	0.0462
	m_3C_3	4	0.0615
	m_4C_3	9	0.1385
	m_5C_3	7	0.1077
		Total 65	0.9696

Figure 3 below shows a graph of chronological change of JHD entropy values from one version to another. To construct the graphs displayed in figure 3, entropy calculated for a version was compared to the previous one. As depicted, it should be noted that initially, the entropy remains

stubbornly the same, but at a later stage, the entropy dropped consistently as the test software versions transition from one version to another.

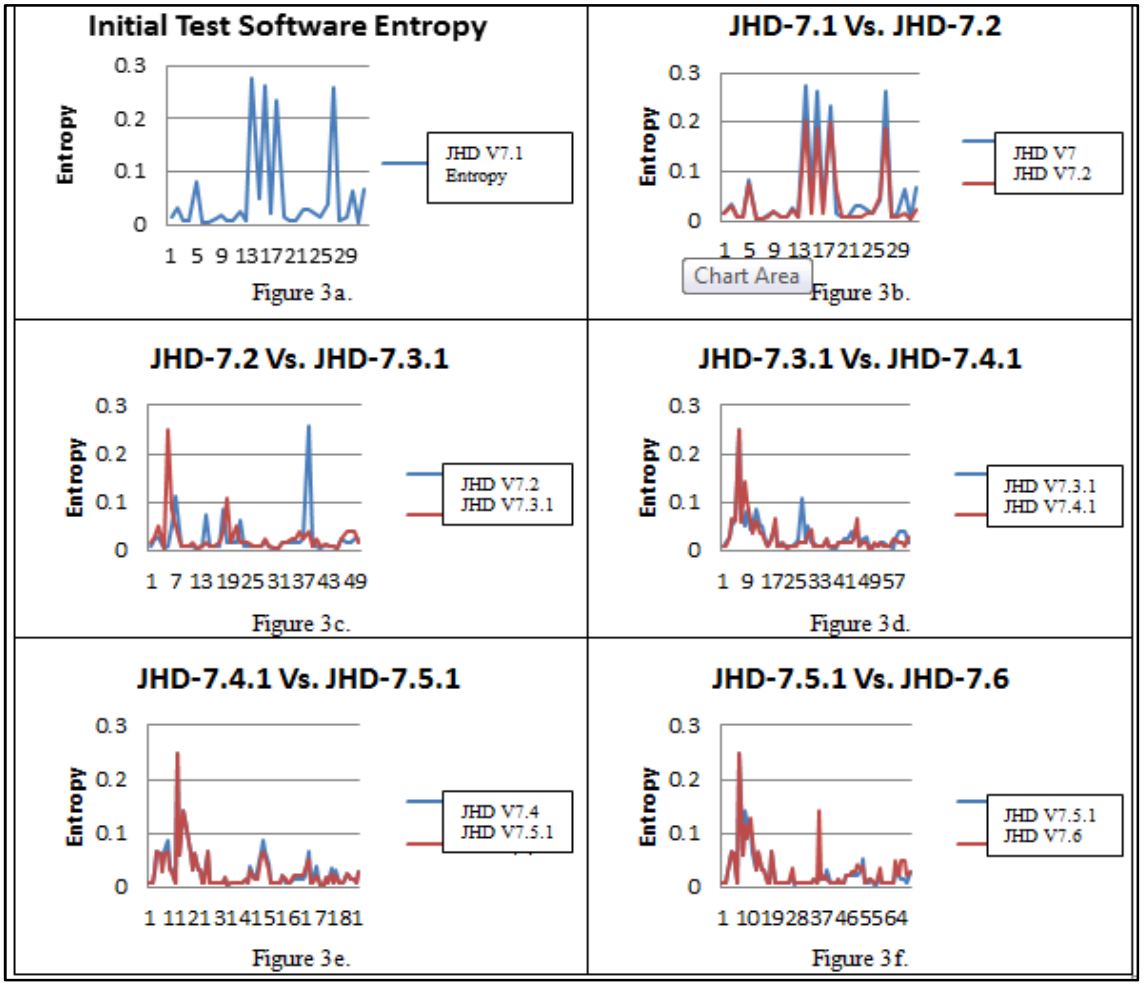

Figure 3 Entropy graph Version to Version

The graph shown in figure 3a is for the initial version of JHD (version 7.0.1) before any change is made. The subsequent figures (3b through 3f) are a superimposition of entropy values representing transitions from one version to another (two versions at a time). From these graphs, a gradual decrease in entropy values can be observed. The high spikes in the middle of each graph are indications of changes reused packets and other add-in modules have undergone throughout the transitional evolution of the test software system.

4.3.2 Software Maturity Index (SMI)

When discussing software maturity, [6] defined Software Maturity Index (SMI) as a metric that provides an indication of the stability of a software product (based on changes that occur for each release of the product). The software maturity index is computed in the following manner:

$SMI = [M_T - (F_a + F_c + F_d)]/M_T$

Where,

M_T= number of modules in the current release

F_c= number of modules in the current release that have been changed

F_a= number of modules in the current release that have been added

F_d= number of modules from the preceding release that were deleted in current release

Software maturity index (SMI) is especially used for assessing release readiness when changes, additions or deletions are made to an existing software system. An observation made by [6] emphasized that, as SMI approaches 1.0, the product begins to stabilize. SMI may also be used as metric for planning software maintenance activities. The mean time to produce a release of a software product can be correlated with SMI, and empirical models for maintenance effort can be developed. In this project, this metric was derived from the chronology of JHotDraw Updates/Additions/Deletions documented and presented by [5]. In this project, the calculation of SMI is based on the package rather than at class or method granularity levels.

Table 3. Data for Software maturity index calculation

From Version to Version	No. Of Packages	Packages Added	Packages Changed	Packages Deleted	Calculated (SMI)
JHD-V7.1 to JHS-V7.2	46	8	24	0	0.30
JHD-V7.2 to JHS-V7.3.1	46	0	23	0	0.50
JHD-V7.3.1 to JHS-V7.4.1	44	6	0	2	0.81
JHD-V7.4.1 to JHS-V7.5.1	46	3	6	0	0.80
JHD-V7.5.1 to JHS-V7.6	45	1	7	1	0.80

From archive data obtained from [18] and [19], a summary of all addition, changes, and deletions made to JHD versions 7.1 through version 7.6 were used to calculate the software maturity index as shown in table 3 above. Also, from this data, the SMI graph is drawn and displayed in figure 4 below. From this graph, it will be seen that the Maturity Index (MI) increases and then levels off as the optimal level of 0.8 is reached, starting from the evolution transition point (V7.3.1 to V7.4.4), stagnating all the way through (V7.6).

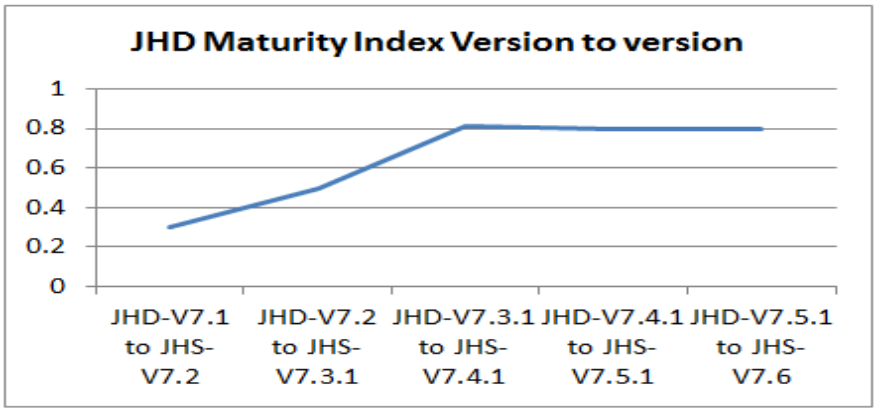

Figure 4 Inter-version Maturity Index

To further view the nature of the JHD evolution and the attained maturity pictorially, the SMI is calculated from the collected transition data for all versions and graphed as shown in figure 5 below.

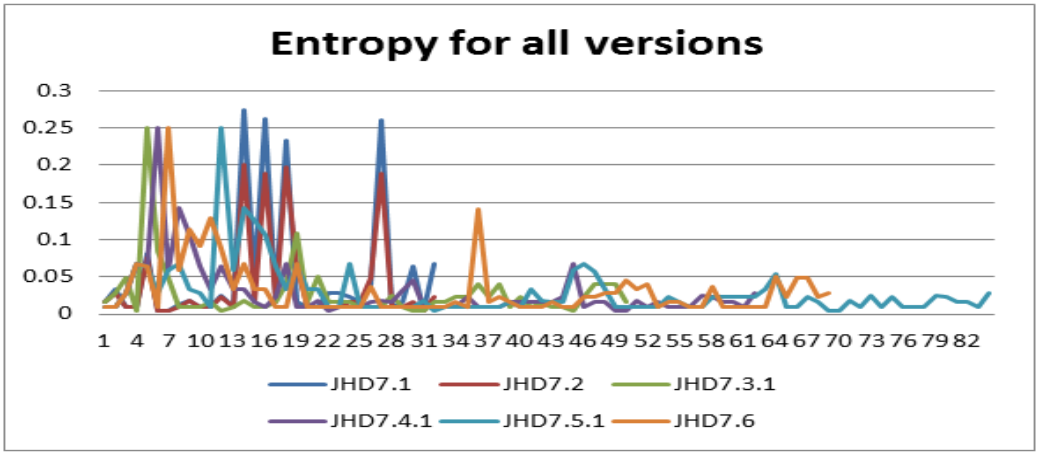

Figure 5. Entropy Values for all 6 versions of JHD

5. ANALYSIS OF RESULTS

On close observation, it will be noted that all versions started with high entropy values, but as soon as the software transitions from JHD7.31 to JHD7.4.1, the entropy starts to drop and then stays consistently at a lower level. If we observe figure 4 above, we can also see that JHD attained its maturity during the transition from JHD 7.3.1 to V.7.4.1. According to [6], a software product reaches its maturity when software maturity index approaches 1. From both figure 6 and 7, we can theorize that in well-designed software that is based on best practices such as JHotDraw, the maturity level is reached at the turning point at which observed entropy starts to decrease. To allow us to visualize the adjustments JHD went through as it transitions and matures. The chart shown in figure 6 are constructed from data extracted from table 1.

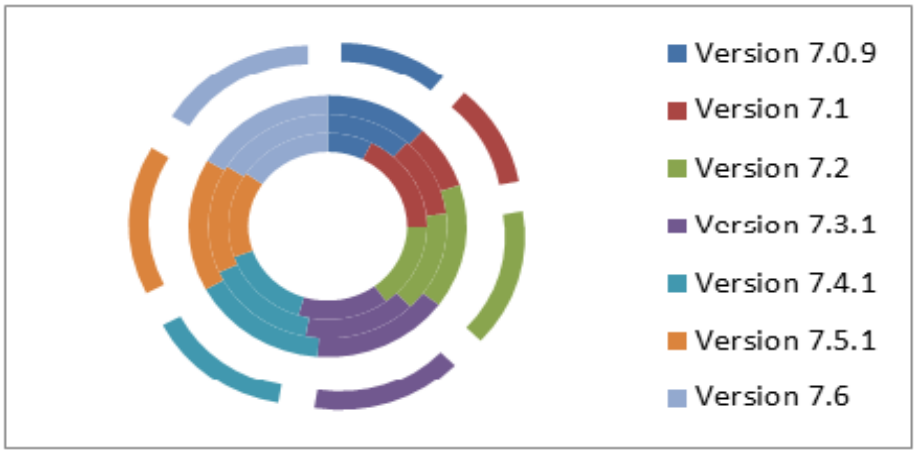

Figure 6. Blown-up Pie Chart for all the six Versions of JHD

It can be seen from the blown-up pie chart shown in figure 6 above that, when initially transitioning from (version 7.09 to 7.1 and from version 7.1 to 7.2), the pie parts in this transition did not align up properly with the outer pie pieces; however, as JHD evolves and transitions (clockwise), the pie pieces started to form perfect alliance with their respective outer pieces, indicating that maturity level has been attained, and the SMI remaining constant at 0.8 for (Version 7.4.1, Version 7.5.1 and the final Version 7.6).

When this observation is compared with the values of maturity index calculated from the static data collection, (see graph in figure 4 above), there is a correlation between the two results, in the sense that the expected maturity level is attained when JHD transitioned from (version 7.4.1 to 7.51); which is the point at which lowest entropy was reached and the highest software maturity index was attained. Another important observation is that, when JHD version transition static data (size, the number of classes, the number of methods and number of attributes) were graphed as shown in figure 7 below, it was observed that the number of methods consistently decreases as the software evolves and transitions from one version to another.

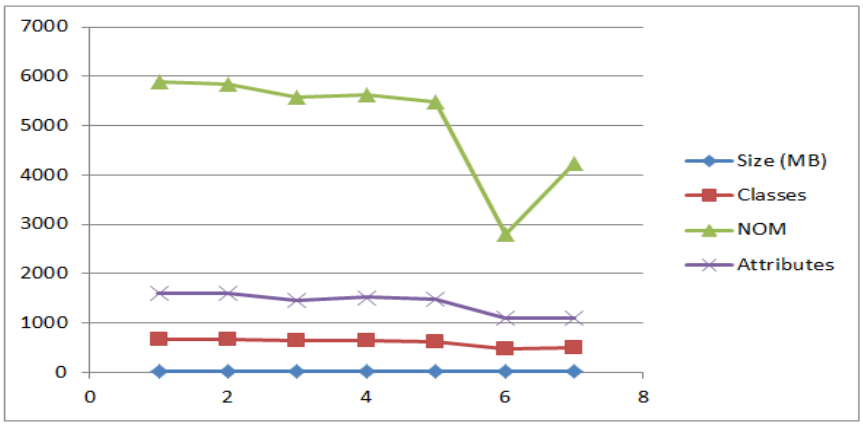

Figure 7. Correlation Between software size, number of classes, methods, and attributes

6. CONCLUSION

When a software system evolves and transitions from one version to another, it is expected that the new version will outperform the previous one and that the new version is better structurally containing fewer defects; however, this may not be the case, as new unintended consequences may be introduced, structure may be degraded and a measure of degradation and disorder may be introduced. This study is a first step towards investigating the behavior of a large-scale matured software system with a view to learn some lessons that can be used as a guideline in design, development, and management of new and existing software systems. In this work, it was consistently observed that JHD software components (classes, methods, and packages) that have undergone change or modifications in JHD evolution tend to generate higher entropy values than those with little or no unchanged; which is in line with an observation by [21] that, the most frequently invoked classes/methods in object-oriented software system are the ones that have the highest possibilities of being changed or modified. It is also observed that the entropy values consistently decreases as the software system evolves from one version to another, indicating that the software system was moving towards its optimal maturity level.

When JHD evolved few versions away from the last version, it is observed that the maximum maturity index attained was (0.8), confirming the statement made by [6] that, a software product reaches its optimal maturity level when its maturity index approaches the value of 1.0. In this research, when the optimal value of 0.8 SMI was reached, the entropy value remains stagnant with little or no change. Also, it was at this turning point that the JHD entropy level tends towards its lowest level, implying a possible correlation or connection between SMI and decrease in entropy, (i.e. decrease in degradation or disorder). In future efforts, we intend to study large-scale, middle-size and small-size object-oriented software systems that have gone through many versions with a view to finding some other hints that may generally be used as a maturity indicator, and a decision guideline for release readiness of software systems.

REFERENCES

[1] Martin and McClure, (1993). Software Maintenance: The Problems and Its Solutions Prentice Hall Professional Technical Reference 1983 ISBN:0138223610

[2] Alessandro Murgia1, A., Concas1, Pinna1, S., Tonelli1, R., Turnu, I. (2009). Empirical, Study of Software Quality Evolution in Open Source Projects Using Agile Practices

[3] Olague, H.M., Etzkorn, L.H., Cox, G.W. (2006). An Entropy-Based Approach to Assessing Object-Oriented Software Maintainability and Degradation-A Method and Case Study. ;In Software Engineering Research and Practice(2006)442-452

[4] Bianchi, A., Caivano, D., Lanubile, F., Visaggio, G. (2001). Evaluating Software Degradation through Entropy, Dipartimento di Informatica - Universith di Bari, Italy

[5] Jones, C. (2006). The economics of Software Maintenance in the Twenty-First Century Version 3 – February 14, 2006. http://www.compaid.com/caiinternet/ezine/capersjones-maintenance.pdf

[6] Pressman, R, Software Engineering - A Practitioner's Approach (6th Ed.). Newyork, NY: McGraw-Hill. p. 679.ISBN 0-07-285318-2

[7] Basili, R. and Rombach, H. D. (1988). The TAME project: Towards Improvement-Oriented Software Environments, IEEE Trans. on Software Engineering SE-14(6) (1988) pp.758–773.

[8] Becker-Kornstaedt, U., and Webby, R. (1999.) A Comprehensive Schema Integrating Software Process Modelling and Software Measurement, Fraunhofer IESE-Report No. 047.99 (Ed.: Fraunhofer IESE, 1999), http://www.iese.fhg.de/Publications/Iese_reports/.

[9] Bhattacharya, P., Iliofotou, M., Neamtiu, I., Faloutsos, M. (2012). Graph-Based Analysis and Prediction for Software Evolution Proceeding of the 34th International Conference on Software Engineering pp. 419-429

[10] Gao, Q., Li, J., Xiong, Y. et al. (2016). High-confidence software evolution. Sci. China Inf. Sci. (2016) 59: 071101. doi:10.1007/s11432-016-5572-2

[11] German, D. (2008). The Past, Present, and Future of Software Evolution. Frontiers of Software Maintenance, FOSM 2008.

[12] Bergmann R. and Eisenecker U. (1996). Case-based Reasoning for Supporting Reuse of Object-Oriented software: A case study, in Proc. Expert Systems 95 (1996) 152–169.

[13] Johari, K., and Kaur, A.(). Effect of Software Evolution on Software Metrics: An Open Source Case Study. ACM SIGSOFT Software Engineering Notes Page 1 September, Volume 36 Number 5, 2011.

[14] Deitel, H.M. & Deitel, P.J., (2003), Java How to Program, Prentice Hall, Upper Saddle River, NN, USA (1

[15] Giesecke (2006). Dagstuhl Seminar Proceedings 06301 Duplication, Redundancy, and Similarity in Software

[16] Mens, Kim., Kellens A., Tonella, P (2007), A Survey of Automated Code-level Aspect Mining Techniques, Transactions on Aspect-Oriented Software Development, Special issue on Software Evolution

[17] http://www.randelshofer.ch/oop/jhotdraw/index.html

[18] http://www.randelshofer.ch/oop/jhotdraw/Documentation/changes.html

[19] Hector M. Olague, Letha H. Etzkorn, Wei Li, Glenn W. Cox (2005). Evolution in software systems: foundations of the SPE classification scheme. Special Issue: IEEE International Conference on Software Maintenance (ICSM2005) Issue Overviews

[20] Opensource Software, www.sourceforge.net

[21] Joshi, P. & Joshi, R. (2006) Microscopic Coupling Metrics for Refactoring, IEEE Conference on Software Maintenance and Software Reengineering.

AUTHOR

Sayyed G. Maisikeli is currently an Assistant Professor at Al-Imam Muhammad ibn Saud Islamic University in Riyadh, Kingdom of Saudi Arabia. He obtained his dual Master of Science degrees in Computer Science and Operations Research from Bowling Green State University Ohio, and his Ph.D. from Nova Southeastern University in Florida, USA. His research interest includes Software evolution, Software visualization, Aspect mining, Software re-engineering and refactoring and Web Analytics.

INTENTIONAL BLANK

A LITERATURE REVIEW ON SEMANTIC WEB – UNDERSTANDING THE PIONEERS' PERSPECTIVE

Salih Ismail[1] and Talal Shaikh[2]

[1]Mathematical and Computer Sciences, Heriot Watt University, Dubai, UAE
si8@hw.ac.uk
[2]Mathematical and Computer Sciences, Heriot Watt University, Dubai, UAE
t.a.g.shaikh@hw.ac.uk

ABSTRACT

There are various definitions, view and explanations about Semantic Web, its usage and its underlying architecture. However, the various flavours of explanations seem to have swayed way off-topic to the real purpose of Semantic Web. In this paper, we try to review the literature of Semantic Web based on the original views of the pioneers of Semantic Web which includes, Sir Tim Berners-Lee, Dean Allemang, Ora Lassila and James Hendler. Understanding the vision of the pioneers of any technology is cornerstone to the development. We have broken down Semantic Web into two approaches which allows us to reason with why Semantic Web is not mainstream.

KEYWORDS

Semantic Web, Literature Review, Pioneers' Perspective

1. INTRODUCTION

This "The power of the Web is in its universality. Access by everyone regardless of disability is an essential aspect." – Sir Tim Berners-Lee (Inventor of the World Wide Web) [1]

The World Wide Web (WWW) was created by Sir Tim Berners-Lee with the vision of connecting people. It didn't take long for WWW to become a global phenomenon and become the backbone for communication on a global level. So much information has been uploaded on the Web that, information could be found just about anything on the Web. This phenomenon kept on growing and essentially the Web has become the brain of planet Earth. [2] There is approximately 100 petabytes of data available on the Internet. [3]

How can you find the right information about a particular topic from this vast ocean of data depends on who wants to find the right information, whether it is a human being or a machine. Search engines have become so important because it helps us retrieve various information about a particular topic. But human beings are capable to compare the different webpages and make an informed choice because, we understand the 'meaning of the information'.

Dhinaharan Nagamalai et al. (Eds) : ICCSEA, SPPR, UBIC - 2016
pp. 15–28, 2016. © CS & IT-CSCP 2016 DOI : 10.5121/csit.2016.61102

The existing Web does provide a fair degree of aid for machines to find information. But definitely that's not enough to find the right information. Currently the extent of the capability of machines are to find the information based on a keyword match and its variants. But machines for a fact do not understand the meaning of those keywords. Many Artificial Intelligence researchers believe that Machine Learning is the key to unlocking machine understanding. However, Machine Learning is hugely based on mathematical equations and statistical analysis. It is computationally very costly even to perform the smallest of tasks, like identifying an image. [4]

"If I have a virtual personal assistant and have somebody who is helping me do my shopping, you are essentially selling to the machine or my agent. Suddenly, that means you need to be good at data. It means that you need to make sure that you have all your products and all the scripts are described in the data that the machine understands." – Sir Tim Berners-Lee [5]

Sir Tim Berners-Lee believes that this is only possible with the adoption of the 'Semantic Web'.

Semantic Web is the evolution of the WWW due to the enhancement in other parallel technologies like pervasive computing, sentient computing, internet of things, artificial intelligence etc. Semantic Web tries to classify the data based on different topics and assign meaning to it. This would in turn not only aid in better human understanding, but also in enhancing the understanding of the machines. Truly the power of WWW can become an extension of the human mind's capacity at this point.

Even though the importance of Semantic Web has been stressed upon by various technology futurists and other respected personalities within the technology world, nothing disruptive has been happening towards this area. [6] [7]

There are tasks and processes which human beings are far superior to computer and vice-versa. Semantic Web would enable machines and humans to cooperatively perform tasks, which would utilize the strength of both realms and provide with a better result.

If machines are able to find commonalities and anomalies among various ontologies, they would be able to create a knowledge base that would really enhance the understanding of the machines. And if machines understand better, then humans would be able to delegate a lot of tasks that a machine could do better and faster. The real essence of 'Co-computing' would become a reality by the use Semantic Web.

2. STRUCTURE OF THE REPORT

The introduction section above, threw light on setting the tone at which the pioneers looked upon at Semantic Web and how do this technology fit in to the realm of the machines.

The section below, deals with the Literature Review, where we explain the basic concepts and terminologies of Semantic Web briefly as per the descriptions and explanations of the pioneers. Finally, we explain few criticisms that Semantic Web faces and what are the replies provided by the creators of Semantic Web.

Finally, we conclude the paper by summarizing the essence of the vision of the creators of Semantic Web and what would be a good start to refresh the perspective about Semantic Web.

3. RELATED WORKS

There is a lot of literature review that has been done on Semantic Web, but very few work has been done reiterating the idea behind Semantic Web from the pioneers' perspective. The real essence of where and why Semantic Web should be used has really deviated. The introduction was a snapshot of the vision of what Semantic Web should be doing. On these lines, Janev and Vrane has spoken about few concepts of what Semantic Web can do and also done a survey on the existing tools and languages available to achieve it [28]. But they are focusing on the constraints of Semantic Web in general and not particular to the bigger role that Semantic Web should be playing.

Guns Raf has done another analysis on tracing back the origins of Semantic Web [29]. He tried to counter debate the criticism of Semantic Web just being Web. They have traced back Semantic Web origins back to early concepts of Artificial Intelligence. This would prove to be in alignment with the concepts of how Semantic Web can be used according to Sir Tim Berners-Lee.

Benslimane et al. discuss the importance of how Semantic Web can become important for machines if there is proper method to structure existing data into Semantic format using RDF and OWL [30]. This is again an outcome of Semantic Web's real use.

We will try to explain the literature that is available to understand Semantic Web in its real essence in the following section.

4. LITERATURE REVIEW

4.1 From Web to Semantic Web

Most people thought WWW wouldn't become successful as there a lot of problems like, who will upload the data, who will manage it, who will fix the issues etc. But the widespread adoption of WWW has been on a planetary level and just about everything has a webpage for it.

The Web infrastructure currently is a distributed network of interlinked webpages with Unique Resource Locators. This helps to categorize webpages of a particular niche and identify them. The idea of Semantic Web is to push the very same infrastructure, where the linking of resources is on the data level. Semantic Web is based on the idea of Smart Data [9].

Figure 1: From Web to Semantic Web.

Smart Data is interlinked data that allows not only humans to use the information, but machines too. Even if each entity of the data is held by individual organization, since they are all interlinked, it could make more meaning [10]. Sir Tim Berners-Lee believed that when interlinked data could also have the property of self-description, it would lead to Semantic Web.

Some of the features of Semantic Web are to be compared with how the Web itself was developed to understand it better:

- The voice – WWW took off to a flying start because anyone was able to say anything about any topic (AAA slogan). There could be hundreds of opinions about a particular topic and it is up to the reader to make a decision. This is one of the striking phenomenon that led to WWW becoming a global endeavor. But this also resulted in the Web becoming a place full of information but hard to find out the "right information". Semantic Web needs to allow the same heterogeneity of data, but at the same time have a small model to start off the discussion on any topic. For this purpose, RDF was created which helps to link data [9].

- Content creators – The most resounding question to the proposition of the Web was "who is going to create pages?" The answer to this question was that "everyone would create content". It proved to be true against the speculation of the skeptics, and to even the proponents of the Web. The same concept needs to be borrowed for Semantic Web.
  The web already proved to us that content wouldn't be a problem and it will eventually be populated. The same goes for Semantic Web. The Web grew because of the 'network effect'. Crowd sourced contents like Wikipedia and IMDB made their entry and grew into massive sources of information [9].

- The users – The Web was meant for humans to share information with one another. Semantic Web has another user – "Machines". One of the major reasons as to why the importance of Semantic Web is increasing is the evolution of Pervasive Computing and Internet of Things. The estimated increase of the number of devices are exponential [7]. This means only one thing – "More Data". If all this data would have been linked to each other, the potential of this data would be massive [9].

4.2 Semantic Modelling

If we adopt the main three principles of the Web for Semantic Web and create non-unique naming, it creates an environment which will allow Semantic Web to grow like a network effect and become a global phenomenon. But a problem still persists; how to find the right plant within a forest?

Finding the right information whether it is linked or non-linked is of utmost importance. Especially Semantic Web is intended for machines too. Humans have the capacity to reason and analyze the 'right information'. But even for humans it would be time consuming to find the correct information from huge collection of data [4].

Even we human beings create an abstraction of information that we come across and then transform it into knowledge. This process of abstraction and the quality of this process is cornerstone to our understanding of a topic. If we are to instil this sort of mechanism to a machine, it is important that there is a model followed.

Modelling is the process of organizing the information. This solves the problem of finding the 'right information' to a great level of accuracy and further provides:

- A framework for human communication.
- Meaning for explaining conclusions.
- Structure for mediating various viewpoints.

If the information is categorized and modelled, it is easier to find the right information. This simple principle is the very reason as to why we have markup languages and frameworks like RDF, OWL etc. They provide a mechanism to model the data and provide semantics [9].

4.3 Resource Description Framework (RDF)

According to W3C "RDF is a standard model for data interchange on the Web. RDF has features that facilitate data merging even if the underlying schemas differ, and it specifically supports the evolution of schemas over time without requiring all the data consumers to be changed." [11]

- An RDF is generally expressed using something known as a triple which is the most basic unit of information.
- A triple contains a subject, a predicate and an object.
- Namespaces are provided to solve the problem on ambiguity in RDF [12].

A simple instance of an RDF document is shown below:

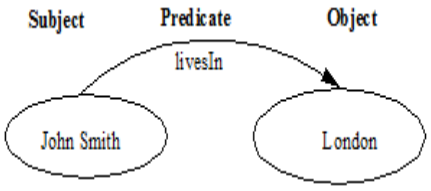

Figure 2: An RDF representation of a fact [13].

The idea of this simple unit of information is that it could be expressed in various formats that could be easily read by machines. The RDFS (RDF Schema) is used for describing the properties and classes of an RDF document. RDFS acts similar to the function of metadata for RDF.

The instance of the triple in the figure is shown in XML and JSON below:

```
<sem:triples xmlns:sem="http://marklogic.com/semantics">
 <sem:triple>
  <sem:subject> http://xmlns.com/foaf/0.1/name/"John Smith"</sem:subject>
  <sem:predicate> http://example.org/livesIn</sem:predicate>
  <sem:object datatype="http://www.w3.org/2001/XMLSchema#string">"London"</sem:object>
 </sem:triple>
</sem:triples>
```

Figure 3: RDF in XML [13].

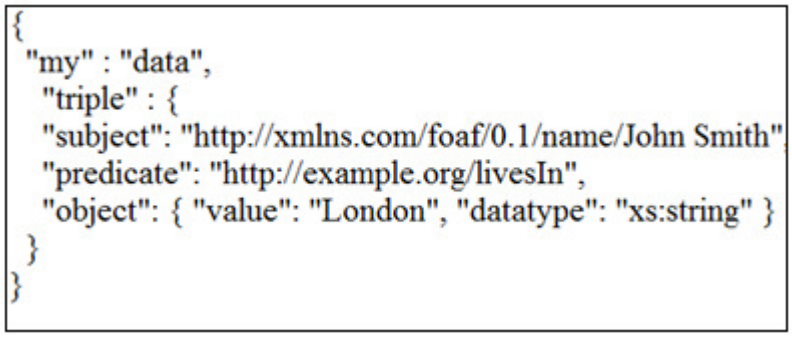

Figure 4: RDF in JSON [13].

4.4 Web Ontology Language (OWL)

Web Ontology Language (OWL) is another standard from the W3C Consortium to aid in the progress of Semantic Web.

OWL provides greater machine interpretations by providing additional vocabulary along with other semantics. OWL adds more vocabulary to describe the RDFS (RDF Schema). This concept is the fundamental idea behind the improvement of machine understandability by the use of OWL [14].

OWL was categorized into three sublanguages to fit the need of the users:

OWL Lite – A simple sublanguage that provides classification hierarchy and constraints. The cardinality can only be values of 0 or 1, thereby restricting and shrinking down the complexities of relations.

OWL DL – The DL stands for 'Description Logics', which is one of the foundational areas for the creation of OWL. OWL DL is for users who wants to achieve the full expressiveness of a topic while ensuring that the computation will finish on a finite set of time.

OWL Full – OWL Full is for users who needs to traverse the entire hierarchy of a subject to its root and even the metadata of the root. It has no computational guarantee as it is quite understandable that this process could be really complex. However, OWL Full pushes to create all the possible meaning of an RDF class [14].

4.5 Ontology

An ontology doesn't have a formally accepted definition. However, a vocabulary and ontology is often used with the same meaning. An ontology can be defined as a set of URIs that makes up meaning for a particular topic [15].

The units that make up an ontology would be a set of RDF along with OWL. There are various ontologies that has been created and frequently used. However most of the ontologies are created by humans and machines have little to no say in this matter.

Examples of few ontologies:

Dublin Core – They have an ontology for metadata of data. Their ontology set includes classes, properties, vocabulary encoding schemes, syntax encoding schemes and collections [16]. All of them have several set of RDFs explaining what the data is about.

Dbpedia – "It is a crowd-sourced community effort to extract structured information from Wikipedia and make this information available on the Web. Dbpedia allows you to ask sophisticated queries against Wikipedia, and to link the different data sets on the Web to Wikipedia data. Furthermore, it might inspire new mechanisms for navigating, linking, and improving the encyclopedia itself." [17]

The above quotation is the official description of Dbpedia from its authors. It would be fair to say that Dbpedia is a Semantic Web version of Wikipedia. The Dbpedia ontology is massive and has 4.2 million instances of objects which include things, person, place, work, organization, and species. This massive amount of data is structured using RDF and OWL. Some of them are linked to other linked-data sources turning Dbpedia into the nucleus of Web of Data as mentioned [18].

There are various other examples of ontologies like FOAF, Good Relations, Music Ontology etc. [15] Now all these ontologies need to be stored somewhere and for that purpose we have Triple Stores.

4.6 Triple Store

A Triple Store is a specific kind of database store for storing and retrieving triples. They are stored in the format of subject, predicate and object. For instance, "Alice knows Bob", "Alice is 15" etc. They are custom built for the purpose of Semantic Web and Linked Data. Similar to any database, the information is retrieved via a query language. A Triple Store has the ability to import and export the required information in RDF format as well [19].

There are a lot of different variants of Triple Stores, some of them are created from scratch and some of them are built on-top of existing SQL and NoSQL databases. Triple Stores are often also called as RDF stores [20].

Few examples of Triple Stores are:

Virtuoso – It is a middleware that supports traditional Relation Database Management Systems (RDBMS) and also has specialized support for RDF document storage and retrieval. It supports multiple protocols and uses a single multi-threaded process. It is also known as Openlink Virtuoso. It provides a SPARQL end point like all the Triple Stores. Virtuoso is well known for its performance in holding huge datasets. For instance, Dbpedia is hosted on a Virtuoso Triple Store [21].

Fuseki – It is a sub project from Apache Jena. It provides an RDF server that can be a Triple Store, which can be administered and managed via REST protocols. It can run as a service on a remote machine, a WAR (Java Web application file) or as a standalone server. Fuseki supports

SPARQL 1.1 and also has added in logging support to keep a close watch of what happens on the triple store.

Fuseki's latest version v2, provides security through Apache Shiro. It adds cryptography and session management to Fuseki [22].

These Triple Stores have an endpoint for SPARQL to query the RDF documents.

There has been a lot of study conducted to find the most optimal Triple Stores. A comparison done between Apache Fuseki, Blazegraph, Sesame and Virtuoso is shown in Table 2.

Table 1: Comparison of Triple Stores [26]

Name	**License**	**Deployment**	**Language**
Apache Fuseki	Apache License 2	Standalone or WAR	Java
Blazegraph	GPLv2 or commercial	Standalone or WAR	Java
Sesame	BSD	WAR	Java
Virtuoso	GPL	Native	C

A performance benchmarking was done by Vladimir Mironov et al. of various Triple Stores. The findings were also a positive addition to the selection of Fuseki [27].

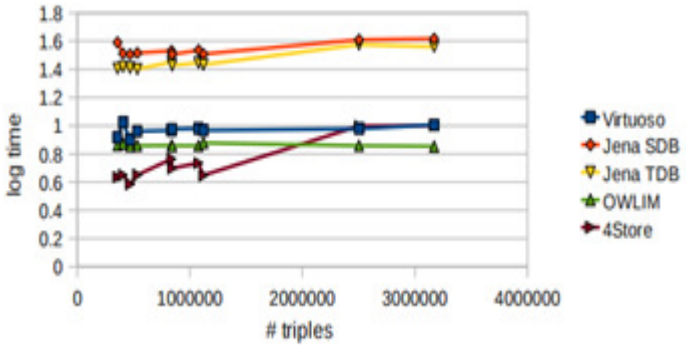

Figure 5: Average Response Time of various Triple Stores [27]

Apache Fuseki is based on Jena TDB and currently only known as Fuseki. So it is evident from the conclusion of the study (See Figure 8), that Fuseki is a winner in case of average response time of queries.

4.7 SPARQL

SPARQL is another W3C standard in the category for Semantic Web. It is a query language similar to that of Structured Query Language (SQL) for Relational Database Management Systems (RDBMS). SPARQL is used to query RDF documents. RDF documents as explained in

Section 2.2.1, are depicted in the format of labelled triples. This allows a graph representation of the RDF document. So SPARQL queries can be result sets of graphs [23].

The results set is either a literal (value) or a URI. The ability to fetch the literal or even convert URI into their labels provide a direct and easy way for applications to use the result set directly.

Table 2: Keywords used by SPARQL [23].

BASE	SELECT	ORDER BY	FROM	GRAPH	STR	isURI
PREFIX	CONSTRUCT	LIMIT	FROM NAMED	OPTIONAL	LANG	isIRI
	DESCRIBE	OFFSET	WHERE	UNION	LANGMATCHES	isLITERAL
	ASK	DISTINCT		FILTER	DATATYPE	REGEX
		REDUCED		a	BOUND	true
					sameTERM	false

Table 2, shows the list of most commonly used keywords used by SPARQL. A SPARQL Abstract Query is a tuple (E, DS, QF) where:

- E is a SPARQL algebra expression
- DS is an RDF Dataset
- QF is a query form

Every triple store has a SPARQL endpoint as mentioned. But triple stores generally can restrict the kind of SPARQL queries than can be executed. So permissions can be controlled at the application level.

5. DISCUSSIONS

We start our discussion with two main views on Semantic Web by different authors on this topic. These approaches can be collated in to two different sets of a Venn Diagram namely, Semantic and Web. The main driver in semantic is artificial intelligence and in Web is Smart Linked Data.

Semantic Approach

As discussed earlier, when Tim Berners Lee spoke about having personal agents, it meant that these software agents would be able to interpret data and accomplish personal task for us. The way this would be done would be with the concept of inference or reasoning. This is a very common approaches in Knowledge based AI where new facts are inferred from existing facts and reasoning can be done on existing datasets. This would give the ability to agents to not only use the information that has been given to them at the start but also create new intelligence on their own.

A survey was done by Jorge Cordoso titled “The Semantic Web Vision: Where are We?”, which has a criterion that is quite interesting to our paper [31]. They list out the top reasons as to why Semantic Web is in use in Figure 5, and sharing common understanding of information structure among people and agents are on top.

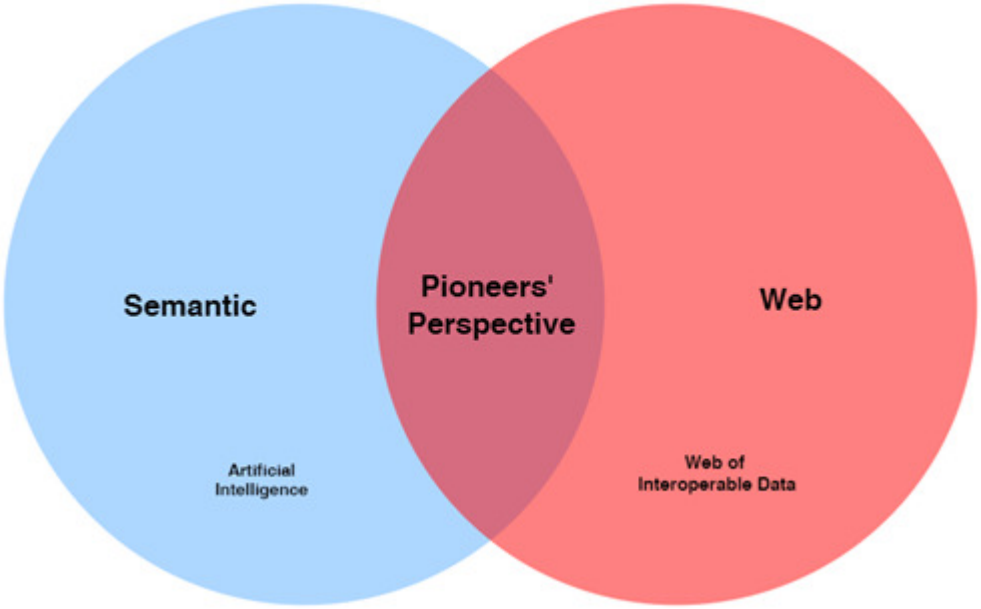

Figure 6: The Pioneers' Perspective

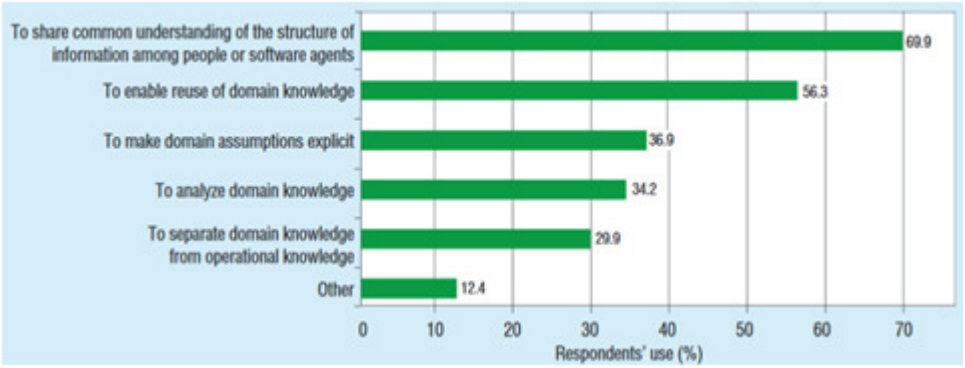

Figure 7: Reason of Semantic Web usage [31].

Roughly about 70% of the users find that Semantic Web's main purpose is to share common understanding of the structure of information. This enhances the machines to reason and infer on the same base knowledge.

The supporters of this approach were interested in the semantic aspects rather than the hyperlinking of resources aspect. In a nutshell, here semantic web is used for performing artificial intelligence research and developing practical solutions.

Web Approach

In this approach, the Linked Data is the driving force. Here the usage of Semantic Web has been to connect data available at different sources. The source could be structured or unstructured giving rise to a flexible model of data usage. For instance, a crowd sourced project called as The Linked Open Data (LOD) has 31+ billion facts in the LOD cloud as of 2014.

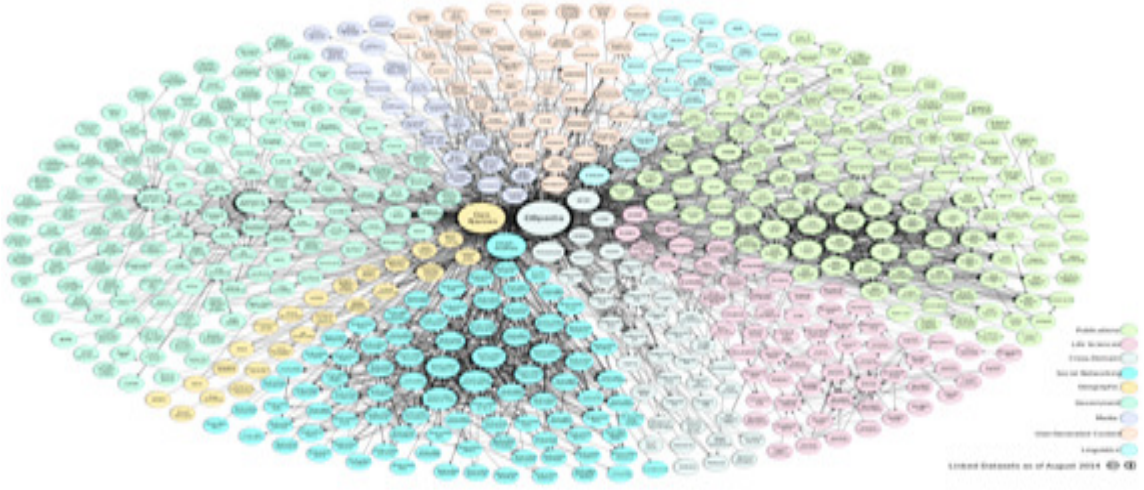

Figure 8 : Linked Open Data Cloud [24]

The supporters of this approach believe the Artificial intelligence is an unwanted liability as they perceive their application would be become more complex.

Semantic and Web Intersection: The Pioneers Approach

In this approach, the pioneers' envisioned the use of Semantic Web where, inferencing is a major aspect. Inferencing and reasoning would essentially lead to knowledge management. This comes from the first part of the semantic approach. But the knowledge that is derived from linked data is more insightful as the data is processed from various sources. And this is based on the second approach, which is the Web approach. But there are quite few concerns shown by the skeptics mainly about the usage of such a system. This is being asked over and over again [25].

James Hendler along with Sir Tim Berners-Lee, responds to this in his presentation by arguing that once we have enough semantic data everyone would want to become a Semantic Web user including governments. The open data project which more and more governments are joining is a clear indication towards this phenomenon [8].

Sir Tim Berners-Lee in his famous TED talk says "The power to ask questions, questions that bridge across different disciplines is a complete sea change." [4]

This is done through Semantic Web. The relationship among various things form the 'bridges' that Sir Tim Berners-Lee speaks about. This is the very basis of enabling cross discipline analysis and research.

James Hendler talks about the application of Semantic Web and how pervasive it has become. He provides enough examples to prove that knowingly or unknowingly everyone uses the fruits of Semantic Web. Facebook's open graph protocol, Oracle's Semantic Web extensions, Google's search result etc. are all ways in which everyone is already a user of Semantic Web [10].

As discussed above we find that these approaches are not mutually exclusive but in practice this is what the pioneers approach actually should be.

6. CONCLUSION

After a brief analysis of the Literature Review in Semantic Web, we are able to understand the vision of the pioneers of Semantic Web. The main target was to transform the current Web to that which has smarter data. This would in turn allow the machines to understand and use the data better.

We have broken down Semantic Web into two approaches and explained them individually to reach to an intersection. Semantic Web is a technology that has great potential for the betterment of the society. But as explained, the focus of researchers is usually only on one approach. And the real value of Semantic Web is at its intersection between the two approaches – The Pioneers' Perspective.

REFERENCES

[1] S. T. Berners-Lee, 2016. [Online]. Available: https://www.w3.org/standards/webdesign/accessibility. [Accessed 05 January 2016].

[2] F. Heylighen and J. Bollen, "The World-Wide Web as a Super-Brain: from metaphor to model," Cybernetics and Systems '96, vol. R. Trappl (ed.), 1996.

[3] C. ROBERTS, 2013. [Online]. Available: http://www.technologybloggers.org/science/how-many-human-brains-would-it-take-to-store-the-internet/. [Accessed 08 January 2016].

[4] S. T. Berners-Lee, "The next Web," TED Talks (https://www.ted.com/talks/tim_berners_lee_on_the_next_web?language=en), 2009.

[5] M. S. GUHA R, "Schema.org: Evolution of Structured Data on the Web," Communications Of The ACM [serial on the Internet], pp. 44-51, 2016.

[6] P. A. Halevy, "The Unreasonable Effectiveness of Data," IEEE Intelligent Systems, vol. 24, no. 2, pp. 8-12, 2009.

[7] statista.com, "Internet of Things (IoT): number of connected devices worldwide from 2012 to 2020 (in billions)," 2016. [Online]. Available: http://www.statista.com/statistics/471264/iot-number-of-connected-devices-worldwide/. [Accessed 20 February 2016].

[8] J. Hendler, Writer, "Why the Semantic Web will Never Work". [Performance]. Keynote, European Semantic Web Conference, Heraklion, June, 2011.

[9] J. Hendler and D. Allemang, Semantic Web for the Working Ontologist, 2nd ed., Elsevier Inc., 2011.

[10] J. Hendler, "Semantic Web: The Insider Story," 2014. [Online]. Available: http://www.slideshare.net/jahendler/semantic-web-the-inside-story. [Accessed 15 March 2016].

[11] W3C, 2016. [Online]. Available: https://www.w3.org/RDF/. [Accessed 15 January 2016].

[12] B. DuCharme, Learning SPARQL, O'Reilly Media Inc., 2013.

[13] M. Community, 2016. [Online]. Available: https://docs.marklogic.com/media/apidoc/8.0/guide/semantics/intro/intro-2.gif. [Accessed 20 January 2016].

[14] W. Consortium, 2004. [Online]. Available: https://www.w3.org/TR/2004/REC-owl-features-20040210/#s1.2. [Accessed 20 January 2016].

[15] Semanticweb.org, 2012. [Online]. Available: http://semanticweb.org/wiki/Ontology. [Accessed 25 January 2016].

[16] D. C. M. Registry, "Browse the registry by classification type," 2016. [Online]. Available: http://dcmi.kc.tsukuba.ac.jp/dcregistry/navigateServlet. [Accessed 20 March 2016].

[17] "Dbpedia - About," Dbpedia.org, 2016. [Online]. Available: http://dbpedia.org/about. [Accessed 15 March 2016].

[18] Dbpedia.org, "Dbpedia.org | Nucleus for the Web of Data," 2016. [Online]. Available: http://wiki.dbpedia.org/use-cases/nucleus-web-data. [Accessed 26 February 2016].

[19] J. Rusher, "Triple Store," 2016. [Online]. Available: https://www.w3.org/2001/sw/Europe/events/20031113-storage/positions/rusher.html. [Accessed 27 February 2016].

[20] K. Cagle, "Semantics + Search : MarkLogic 7 Gets RDF," 22 December 2013. [Online]. Available: http://blogs.avalonconsult.com/blog/search/semantics-search-marklogic-7-gets-rdf/. [Accessed 27 February 2016].

[21] W3C, "OpenLink Virtuoso," 11 February 2012. [Online]. Available: https://www.w3.org/2001/sw/wiki/OpenLink_Virtuoso. [Accessed 18 March 2016].

[22] jena.apache.org, "Apache Jena Fuseki," 2015. [Online]. Available: https://jena.apache.org/documentation/fuseki2/index.html. [Accessed 10 March 2016].

[23] W3C, "SPARQL Query Language for RDF," 2008. [Online]. Available: https://www.w3.org/TR/rdf-sparql-query/. [Accessed 27 February 2016].

[24] U. o. Manheim, "State of the LOD Cloud 2014," 30 August 2014. [Online]. Available: http://linkeddatacatalog.dws.informatik.uni-mannheim.de/state/. [Accessed 06 March 2016].

[25] M. Bernstein, "MIT CSAIL Research," October 2009. [Online]. Available: http://haystack.csail.mit.edu/blog/2009/10/25/tales-of-a-semantic-web-skeptic/. [Accessed 5 March 2016].

[26] utecht.github.io, "Comparing Triple Stores," UTecht Blog, 2015. [Online]. Available: http://utecht.github.io/comparing-triplestores/. [Accessed 12 March 2016].

[27] V. Mironov, N. Seethappan, W. Blondé, E. Antezana, B. Lindi and M. Kuiper, "Benchmarking triple stores with biological data," Proceedings of the 3rd International Workshop on Semantic Web Applications and Tools for the Life Sciences, December, 2010.

[28] V. Janev and S. Vranes, “Semantic Web Technologies: Ready for Adoption?,” IT Professional, vol. 11, no. 5, 2009.

[29] G. Raf, “Tracing the origins of the semantic web,” Journal of the American Society for Information Science and Technology, vol. 64, no. 10, pp. 2173-2181, 2013.

[30] S. M. Benslimane, M. Malki and A. Lehirech, “Towards ontology-based semantic web from data-intensive web: A reverse engineering approach,” IEEE International Conference on Computer Systems and Applications, pp. 771-778, 2006.

[31] J. Cordoso, “The Semantic Web Vision: Where Are We?,” IEEE Intelligent Systems, vol. 22, no. 5, 2006.

AUTHORS

Salih Ismail received bachelors (hons) degree in Information Technology from University of Bedfordshire. He is currently completed in MSc in Computer Systems Management from Heriot Watt University. He has worked on finding commonalities between Semantic entities. His research areas of interest include Network Security, Semantic Web, Internet of Things etc. He has been running an IT company in Dubai.

Talal Shaikh is currently Assistant Professor at Heriot Watt University Dubai in the School of Mathematics and Computer Science (MACS). Some of the subjects he teaches are Network Applications, Software Engineering and Artificial Intelligence. His research area of interest is making "things" such as robots and devices "talk" to one another using software in order to carry out a task and achieve a common goal. His ongoing research covers ubiquitous and pervasive computing, Machine to Machine (M2M) technologies and the Internet of Things. His academic area of interest covers student learning and student experience.

FILE SYNCHRONIZATION SYSTEMS SURVEY

Zulqarnain Mehdi[1] and Hani Ragab-Hassen[2]

[1]School of Mathematical and Computer Sciences, Heriot-Watt University
zm82@hw.ac.uk
[2]School of Mathematical and Computer Sciences, Heriot-Watt University
h.ragabhassen@hw.ac.uk

ABSTRACT

Several solutions exist for file storage, sharing, and synchronization. Many of them involve a central server, or a collection of servers, that either store the files, or act as a gateway for them to be shared. Some systems take a decentralized approach, wherein interconnected users form a peer-to-peer (P2P) network, and partake in the sharing process: they share the files they possess with others, and can obtain the files owned by other peers.

In this paper, we survey various technologies, both cloud-based and P2P-based, that users use to synchronize their files across the network, and discuss their strengths and weaknesses.

KEYWORDS

Cloud storage, Peer-to-Peer, P2P, BitTorrent, & Synchronization

1. INTRODUCTION

Sharing digital files over a network is a common application of the networking technology. Files can be shared between: a) Users and machines (eg: when one downloads a file from a server, or uploads a file to a server), b) Machines (eg: automated backups), c) Different users (through machines: uploading the file to a server, from which the other party can download it; directly: using P2P file sharing services). It is common nowadays for users to share files across their own devices connected over a network using synchronization services such as Dropbox [1] or Google Drive [2]. They usually do so by allowing a user to upload their files from one device to central servers, and allowing other devices owned by the same user to download them from those servers. Note that the users also have the option to share their files with others, or make them public. Peer-to-peer (P2P) based synchronization systems split the files into chunks (or pieces), which are then replicated on a subset of peers. The original peer's other devices can retrieve the chunks and combine them to form the original file. In this paper we review the main existing file synchronization systems, and compare them. The rest of this paper is organized as follows: section II introduces important backbone technologies. Notable existing file synchronization systems are reviewed in section III. We discuss those system in section IV. Section V concludes our paper

Dhinaharan Nagamalai et al. (Eds) : ICCSEA, SPPR, UBIC - 2016
pp. 29–38, 2016. © CS & IT-CSCP 2016 DOI : 10.5121/csit.2016.61103

2. BACKBONE TECHNOLOGIES

Prior to diving deeper into the survey, it is important to have some knowledge of the underlying technologies of the reviewed services. While the technologies used by cloud-based storage systems are quite straight-forward, P2P infrastructures are of a more complex nature. This is why we briefly describe important P2P protocols in this section.

It should be noted that P2P might have slightly different meanings in different contexts. The definition presented in this paper is according to [3]. In a P2P system, each peer in the system provides the service that it is intended to, by sharing its resources (eg: storage and processing power). These peers communicate directly, without the need of an intermediate node.

A pure P2P system is fully decentralized, but partially centralized P2P systems do exist. The best example of such a P2P system is the BitTorrent protocol, wherein a central server, a tracker, tracks the peers currently downloading each file. Other peers can then contact the tracker and request the list of these peers, and contact them.

For a truly decentralized P2P network to exist, the nodes first need to find other nodes in the network (peer discovery). In a local network, a simple scan could reveal other nodes that participate, or are interested in participating in a local P2P network. Over a wider network (such as the Internet), however, this would be a non-trivial task, as it would be unfeasible for a node to look up the entire network to connect to the ones that share the same interests.

There are a few protocols that allow peers to discover each other in a P2P network. In the following subsection, we review Pastry [4], a P2P discovery protocol; we then review BitTorrent, the defacto P2P standard.

2.1. Pastry

In [4], Rowstron and Druschel presented Pastry a new object location protocol for large scale P2P systems. Pastry performs application level node look up and routing over a large network connected via the Internet; when a node receives a message along with a key, it routes the message along all the live nodes to the node which has a nodeID 'numerically' closest to the key. Each node in Pastry keeps track of its immediate neighbors, and notifies other nodes of any changes in the network, such as when a new node joins the network, or if one leaves the network.

Pastry is completely decentralized, and aims to reduce the routing steps that messages have to take to reach the destination. The expected number of routing steps in Pastry is O(log N), where N is the number of Pastry nodes in the network.

A nodeID is randomly calculated, and ranges from 0 to 2128−1, allowing them to be "diverse in geography, ownership, jurisdiction, etc." A node is said to be "close" to be another node if its nodeID is numerically close to the key that it receives along with the message. The message is routed to one of such closest nodes in Pastry, which is usually a node near the originator node.

An example of an application of Pastry is PAST [5], a largescale P2P file storage utility, developed by the same authors. More on PAST is detailed in the next section.

2.2. BitTorrent

According to the official specification, [6], BitTorrent is a P2P file sharing protocol, used to transfer files of any size across the web, and according to [7], it was created by Bram Cohen to replace the standard FTP. It uses a server (called a tracker) that tracks the files, and aids the clients in downloading and combining the chunks (pieces, according to [6]) of the file into the original file. There are, however, "trackerless" implementations of the protocol, which create a true decentralized environment for BitTorrent based P2P file transfers.

Unlike a typical P2P network, BitTorrent ensures that each client uploads files while downloading files from other peers, ensuring fairness, better availability of files, and a boost in performance.

3. FILE SYNCHRONIZATION SYSTEMS

This section presents the most notable file synchronization systems. We distinguish two major categories: cloud-based file synchronization systems, and P2P-based file synchronization systems.

3.1. Cloud-based File Synchronization Systems

A cloud-based synchronization system (also a cloud-based storage service) is used to store users' files in a central server, owned and governed by a certain entity (eg: an enterprise, or a small company). Users upload their files to this server from one device, and download them on another (or on the same device, in case the user loses the original file). Users can also share their files with others, and depending on the service provided, a cloud-based synchronization service can be extended to provide a collaboration platform to the users.

These services are provided across many different platforms, using web as well as native application development technologies as their front-end. Some of them provide desktop applications that act as drives connected to the PC, to provide a seamless interaction with the actual cloud drive. These services usually employ a freemium model: a fixed amount of initial storage is given for free, with limited feature set, while allowing users to upgrade to a higher plan with more storage and additional features. A good comparison of some of the most popular cloud storage and synchronization services can be seen in [8]. Such a model makes cloud services much more accessible and convenient to the users.

3.1.1. Google Drive

Google Drive [2] is a file storage and synchronization service by Google. At the time of writing this paper, new users to the service get 15 GB of storage for free, with various monthly subscription plans available for more storage [9].

Users can not only store and synchronize their files using Google Drive, they can also view, modify, delete, and in some instances, collaborate on them with other users, using either the web interface, or a native applications available on major platforms. Google Drive supports a plethora of file formats for a user to store, synchronize, and work with.

3.1.2. OneDrive

OneDrive [10] by Microsoft is a file storage and synchronization service with similar features to Google Drive, and is powered by Microsoft Azure [11], Microsoft's cloud computing platform. As of January 2016, OneDrive has dropped down its storage capacity for new users from 15 GB to 5 GB. Users who had obtained the 15 GB previously would retain it. Like Google Drive, OneDrive allows users to upgrade the storage using one of the various monthly subscription plans [12].

Along with file storage and synchronization, OneDrive allows users to view, update and delete the files, and collaborate on them using Office Online - a free online Microsoft Office utility.

3.1.3. iCloud Drive

A cloud storage and synchronization service identical to Google Drive and OneDrive, iCloud Drive [13] by Apple offers similar features to users as the previously mentioned cloud-based services. In terms of file storage capacity, iCloud Drive offers 5 GB of free space to new users, like Microsoft's OneDrive, with plans for upgrade available [14].

According to [15], iCloud is utilizes both Amazon Web Services (AWS) by Amazon [16], and Microsoft Azure [11] since 2011 (when iCloud first launched). However, there are numerous reports which state that Apple is siding with Google's Google Cloud Platform [17] to provide some of iCloud's services [18] [19] [20].

3.1.4. Dropbox

Dropbox [1] is one of the most popular file storage and synchronization service, created not by large entities such as those mentioned above, but by a startup company of the same name.

Dropbox offers 2 GB of storage space initially to new users, with options to upgrade to 1 TB, with a monthly subscription (or unlimited storage for Business users) [21].

3.2. P2P-based File Synchronization Systems

A P2P-based synchronization system, unlike a cloud-based synchronization system, is a decentralized system wherein each peer in the network acts as both a server, as well as a client, to synchronize files between a user's authorized devices. In this system, files are broken down into encrypted pieces, and each peer uploads a certain number of pieces to, and downloads from, other nodes in the system, ensuring that the files are almost always available for synchronization, and that no one peer contains the complete file, thus enforcing privacy and security of the users' data. Furthermore, the load is divided among the connected peers, rather than a single server, thus increasing the performance of the synchronization process.

Like centralized cloud synchronization services, P2P service providers provide a similar business model of a free though limited plan, while setting additional storage space up for purchase. However, unlike centralized cloud storage and synchronization services, it is much more efficient and convenient to conjure a private P2P cloud service with possibly unlimited storage (as storage space depends upon the storage shared by each node many nodes equal a lot of storage).

Below are some of the examples of such a system.

3.2.1. PAST

PAST is an application of Pastry, developed by the developers of Pastry themselves. PAST extends Pastry's capabilities to form a peer-to-peer file storage system that uses a file's name, as well as the owner's name, to calculate a hash which is used as its fileID. The fileID is used as the key in PAST.

3.2.2. Symform

Symform by Quantum [22] is a popular P2Pbased file synchronization service, in which each node forms a cloud in the decentralized network, and contributes its resources (storage space), while receiving certain amount of space itself from other nodes.

In Symform, files are broken down into blocks, encrypted, and spread across the network. This way, the files are always available for synchronization, privacy is maintained, security is enforced, and the synchronization performance enhanced on the network.

3.2.3. Resilio Connect

Resilio Connect (formerly Sync, by BitTorrent, Inc.) [23] creates a P2P cloud using BitTorrent among a user's devices, rather than including external nodes into the network. This makes the cloud even more secure, but reduces the reliability of the synchronization service, as offline nodes cannot transmit or receive files.

4. DISCUSSION

Table 1 compares the existing technologies and services we mentioned in the previous section. As can be seen from the table, P2P-based file synchronization systems tend to offer the most value to the consumers than the cloud-based services in terms of storage capacity.

P2P-based systems offer potentially unlimited storage, as each node in the network acts as a server as well as a client. Furthermore, since the pieces of files are replicated on multiple nodes, even if a node is (or a set of nodes containing those pieces are) offline, downloaders can obtain those pieces from the online nodes, thus making the files more readily available for synchronization, and the network more reliable. All nodes in the network need to go offline at the same time for the network to be completely down. Moreover, since the pieces are encrypted, and scattered across the network, security and privacy are ensured in such systems.

Resilio Connect is the only P2P-based synchronization system that is powered by BitTorrent, and inherits almost all the benefits of other P2P-based systems. Although the table shows that Resilio Connect may not have the same level of performance and availability of files as the other systems, a BitTorrent powered synchronization can, in fact, be developed with these advantages.

In what follows, we discuss why we expect Resilio Connect, and more generally BitTorrent-based systems, to be the go-to technology for file synchronization systems.

Table 1. Comparison of existing file synchronization technologies and services.

	Google Drive	OneDrive	iCloud	Dropbox	Symform	Resilio Connect	PAST
Free initial storage (GB)	15	5	5	2	10	Unlimited	Unlimited
Cheapest storage plan	100 GB ($1.99/month)	50 GB ($1.99/month)	50 GB*	1 TB ($9.99/month)	100 GB ($10/month)	Unlimited storage for free	Unlimited for free
Maximum storage	30 TB	1 TB (with Office 365)	1 TB	1 TB	1 TB	Unlimited storage	Unlimited storage
Platform support	Windows, OSX, Android, iOS, Web	Windows, OSX, Android, iOS, Web	Windows, OSX, Android, iOS, Web	Windows, OSX, Linux, Android, iOS, Web	Windows, OSX, Linux, Android, iOS	Windows, OSX, Linux, Android, iOS	Not mentioned(potentially every platform)
Cloud-based or P2P-based	Cloud-based	Cloud-based	Cloud-based	Cloud-based	P2P-based	P2P-based (BitTorrent)	P2P-based
Online collaboration	Yes	Yes	Yes	Yes	No	No	No
File storage before synchronization	Files are stored on the cloud drive (and on the local drive of the device that uploaded the file, if not deleted)	Files are stored on the cloud drive (and on the local drive of the device that uploaded the file, if not deleted)	Files are stored on the cloud drive (and on the local drive of the device that uploaded the file, if not deleted)	Files are stored on the cloud drive (and on the local drive of the device that uploaded the file, if not deleted)	Files are transferred from the device to the P2P cloud formed by Symform, and stored there as long as the user wishes	Files are stored in the devices where they are first added, and transferred only when a sibling device comes online	Files are transferred from the device to the P2P cloud formed by PAST, and stored there as long as the user wishes
File storage after synchronization	Files stay in the cloud drive for as long as the user wishes	Files stay in the cloud drive for as long as the user wishes	Files stay in the cloud drive for as long as the user wishes	Files stay in the cloud drive for as long as the user wishes	The pieces of the files stay in the P2P cloud, as well as the synchronized devices, for as long as the user wishes	The files remain on the synchronized devices, unless the user deletes them	The pieces of the files stay in the P2P cloud, as well as the synchronized devices, for as long as the user wishes
Availability of files	Files are available for download at any time, as long as the cloud hosting them are up and running	Files are available for download at any time, as long as the cloud hosting them are up and running	Files are available for download at any time, as long as the cloud hosting them are up and running	Files are available for download at any time, as long as the cloud hosting them are up and running	Files are available for download at any time, as long as the nodes with the pieces of the files are online	Files are available for download as long as at least two nodes with the files to be synchronized are online	Files are available for download at any time, as long as the nodes with the pieces of the files are online

* Different regions have different prices; US has the cheapest price at $1.99/month

4.1. BitTorrent Advantages

The main reasons in focusing on BitTorrent in this paper to give insights on the superiority of BitTorrent-powered, P2Pbased file sharing and synchronization systems are:

4.1.1. Popularity

According to statistics released by BitTorrent, Inc. [24], there are 45 million daily active users, whereas on a monthly scale, a staggering 170 million users are active each month.

BitTorrent is very popular among the younger population, with 63% of the users aged 34 and below [24]. Furthermore, most of these users are “educated and tech-savvy” males, according to BitTorrent.

It should be noted that, although coming from the official website, these statistics are not complete, as it is rather difficult to collect stats on BitTorrent, due to its nature of being used in a decentralized, and at many times a private networking environment.

4.1.2. Availability

Since BitTorrent is a P2P network, the complete file is almost always available to be downloaded, as long as a single peer is online in the network (assuming it contains the whole file). Furthermore, since the files are divided into pieces, individual pieces can be downloaded from the online nodes. Missing pieces can be downloaded from nodes once they come online. Comparing this to a client/server architecture, wherein a server holds the file to be downloaded, one can definitely see how reliable a P2P network is, more so the BitTorrent protocol.

4.1.3. Performance

Several research works focused on the capabilities of a P2P network, many of which report the performance gains when downloading files using BitTorrent.

Raymond et al. measured the load on centralized servers when using BitTorrent conjointly [25]. The paper showed how BitTorrent reduces the load on a server, and increases the download performance. Using various technologies and measurements, this research presents various tests and analyses results on the performance of the BitTorrent protocol.

As noted above, along with the performance gains, BitTorrent, being a P2P protocol, also reduces the server load by making each node in the network act like a server. Moreover, the network adjusts accordingly to new nodes joining it, or nodes going offline, thus making the network more scalable.

4.1.4. Scalability

In a P2P system, each client is a potential server. That is, increasing demand translates into increasing offer. This results in the unique scalability that characterizes P2P systems. This is unlike a typical server/client architecture, in which a server has to handle an increase, or even a decrease in the number of connected clients. An increase in the number of clients increases the server load, whereas a decrease in the number makes the system less efficient.

4.2. BitTorrent Limitations

BitTorrent may inherit the advantages of a P2P network, but it does come with its limitations. The most prominent limitation of the protocol being is its security. There is a number of well-known security holes in BitTorrent [26], [27], including Authentication, Authorization and Trust & Reputation.

We reviewed the already available P2P file synchronization technologies that have already implemented security in their systems. One of the best examples of such a service is Symform, which encrypts file chunks before replicating them on the network [28]. These systems provide confidentiality and data integrity by encrypting the file chunks. They also provide authentication of the user, by a username and password combination, prior to sharing or downloading files.

5. CONCLUSIONS

We reviewed various file sharing and synchronization technologies and services in this paper. We also compared and discussed these technologies and services, and presented our arguments on why we believe that P2P-based, or more specifically, BitTorrent powered file synchronization systems are superior to traditional cloud-based file synchronization systems, and should be the go-to technologies for reliable and secure file sharing and synchronization services. Future works should focus on enabling online collaboration over P2P-based synchronization systems.

REFERENCES

[1] Dropbox, "Dropbox," http://www.dropbox.com, [Online; accessed 14June-2016].

[2] Google, "Google drive," https://drive.google.com, [Online; accessed 14June-2016].

[3] G. Camarillo, "Peer-to-peer (p2p) architecture: definition, taxonomies, examples, and applicability," 2009.

[4] A. Rowstron and P. Druschel, "Pastry: Scalable, decentralized object location, and routing for large-scale peer-to-peer systems," in Middleware 2001. Springer, 2001, pp. 329–350.

[5] P. Druschel and A. Rowstron, "Past: A large-scale, persistent peerto-peer storage utility," in Hot Topics in Operating Systems, 2001. Proceedings of the Eighth Workshop on. IEEE, 2001, pp. 75–80.

[6] B. Cohen, "The BitTorrent protocol specification version 11031," http://www.bittorrent.org/beps/bep 0003.html, 2013, [Online; accessed 14-June-2016].

[7] J. Fonseca, B. Reza, and L. Fjeldsted, "Bittorrent protocol – btp/1.0," http://jonas.nitro.dk/bittorrent/bittorrent-rfc.html, 2005, [Online; accessed 14-June-2016].

[8] S. Mitroff, "Onedrive, dropbox, google drive and box: Which cloud storage service is right for you?" http://www.cnet.com/how-to/onedrivedropbox-google-drive-and-box-which-cloud-storage-service-is-right-foryou/, 2016, [Online; accessed 14-June-2016].

[9] Google, "Google drive storage plans and pricing," https://support.google.com/drive/answer/2375123?hl=en, [Online; accessed 14-June-2016].

[10] Microsoft, "Onedrive," http://onedrive.live.com, [Online; accessed 14June-2016].

[11] ——, "Microsoft azure: Cloud computing platform & services," https://azure.microsoft.com/en-us/, [Online; accessed 14-June-2016].

[12] ——, "Microsoft onedrive plans," https://onedrive.live.com/about/enUS/plans/, [Online; accessed 14-June-2016].

[13] Apple, "icloud," http://www.icloud.com, [Online; accessed 14-June2016].

[14] ——, "icloud storage plans and pricing," https://support.apple.com/enae/HT201238, [Online; accessed 14-June-2016].

[15] ——, "ios security guide," https://www.apple.com/business/docs/iOS Security Guide.pdf, [Online; accessed 14-June-2016].

[16] Amazon, "Amazon web services - cloud computing services," https://aws.amazon.com/, [Online; accessed 14-June-2016].

[17] Google, "Google cloud computing, hosting services & apis," https://cloud.google.com/, [Online; accessed 14-June-2016].

[18] MacRumors, "Apple inks deal to use google cloud platform for some icloud services," http://www.macrumors.com/2016/03/16/apple-icloudgoogle-cloud-platform/, [Online; accessed 14-June-2016].

[19] CRN, "Cloud makes for strange bedfellows: Apple signs on with google, cuts spending with aws," http://www.crn.com/news/cloud/300080062/cloud-makes-for-strangebedfellows-apple-signs-on-with-google-cuts-spending-with-aws.htm, [Online; accessed 14-June-2016].

[20] B.Insider, "Google nabs apple as a cloud customer," http://www.businessinsider.com/google-nabs-apple-as-a-cloudcustomer-2016-3, [Online; accessed 14-June-2016].

[21] Dropbox, "Dropbox plans comparison," https://www.dropbox.com/business/plans-comparison, [Online; accessed 14-June-2016].

[22] Symform, "Symform: Free online backup service," https://www.symform.com/, [Online; accessed 14-June-2016].

[23] Resilio, "Bittorrent sync," http://www.getsync.com, [Online; accessed 14-June-2016].

[24] BitTorrent, "Bittorrent - advertise with us," http://www.bittorrent.com/lang/en/advertise, [Online; accessed 14June-2016].

[25] R. L. Xia and J. K. Muppala, "A survey of bittorrent performance," Communications Surveys & Tutorials, IEEE, vol. 12, no. 2, pp. 140– 158, 2010.

[26] R. Guha and D. Purandare, "Security issues in bittorrent like p2p streaming systems," SIMULATION SERIES, vol. 38, no. 4, p. 423, 2006.

[27] M. Barcellos, "Security issues and perspectives in p2p systems: from gnutella and bittorrent," http://webhost.laas.fr/TSF/IFIPWG/Workshops&Meetings/53/workshop/ 8.Barcellos.pdf, 2008, [Online; accessed 14-June-2016].

[28] Symform, "The most secure cloud storage — symform," http://www.symform.com/how-it-works/security, [Online; accessed 14-June-2016].

AUTHORS

Zulqarnain Mehdi (Zul) is currently pursuing his MSc degree in IT (Software Systems) from Heriot-Watt University, Dubai. He is currently employed as a Software Engineer in a Dubai-based company.

Zul's research interests include cloud storage systems, file sharing, peer-to-peer, and BitTorrent.

Hani RAGAB received the MSc degree from the university of technology of Compiegne (UTC), France, in 2003, and the Ph.D degree from the same university in 2007. He is currently a lecturer at Heriot-Watt University, United Kingdom.

His research interests include malware analysis, access control systems, peer-to-peer, and digital forensics.

FACIAL EXPRESSION RECOGNITION USING DIGITALISED FACIAL FEATURES BASED ON ACTIVE SHAPE MODEL

Nan Sun[1], Zheng Chen[2] and Richard Day[3]

Institute for Arts, Science & Technology
Glyndwr University
Wrexham, United Kingdom
bruce.n.sun@gmail.com[1]
z.chen@glyndwr.ac.uk[2]
r.day@glyndwr.ac.uk3

ABSTRACT

Facial Expression Recognition is a hot topic in recent years. As artificial intelligent technology is growing rapidly, to communicate with machines, facial expression recognition is essential. The recent feature extraction methods for facial expression recognition are similar to face recognition, and those caused heavy load for calculation. In this paper, Digitalized Facial Features based on Active Shape Model method is used to reduce the computational complexity and extract the most useful information from the facial image. The result shows by using this method the computational complexity is dramatically reduced, and very good performance was obtained compared with other extraction methods.

KEYWORDS

Facial Expression Recognition, Active Shape Model, Feature Digitalisation, Computational Complexity Reduction

1. INTRODUCTION

Computers are behaving more and more likely as human. They can talk and play with a human, but a deeper communication needs interactions of emotions. Thus Facial Expression Recognition (FER) function is important and highly required for future computers.

Since Paul Ekman introduced the 6 basic expressions [1], many FER methods have been developed. Local Binary Pattern (LBP) for texture analysis was introduced by T. Ojala [2, 3]. Then, T. Ahonen presented LBP-based methods for face detection and recognition [4, 5]. X. Tan solved the problem with difficult lighting conditions using LBP [6]. Some researchers believe that using Gabor wavelets for FER is a better way. M. Bartlett did his work utilizing the Gabor filter to recognize the facial expression [7]. The temporal extension of Gabor features in facial expression analysis work was done by L. Ma [8]. Both LBP and Gabor filters deal with the image pixels directly, the extracted features are therefore still related to pixel information. That means the facial images need to be aligned or treated by blocks. The disadvantage of using the above methods is that the complexity of calculation during training and classifying relies on the resolution of the image.

Dhinaharan Nagamalai et al. (Eds) : ICCSEA, SPPR, UBIC - 2016
pp. 39–46, 2016. © CS & IT-CSCP 2016 DOI : 10.5121/csit.2016.61104

Active Shape Model (ASM) was developed by T. Cootes in 1992 [9], which detects the profile of the image. Unlike LBP and Gabor filters, the result of ASM for facial feature extraction is not related to pixels but the positions of facial landmarks. Many variations of the ASM method for FER have been introduced: Optimal Features ASM (OFASM) is high in accuracy but is more computationally expensive [10]. F. Sukno extended OFASM to allow application in more complex geometries [11]. However, the above methods do not consider the wrinkle features, which is quite important in real-world FER.

In this paper, we suggest a new method called Digitalized Facial Features based on Active Shape Model (DFFA) to extract the facial features. ASM is only used to get the landmarks and a part of the useful features. Wrinkle features and other useful facial features are extracted by edge detection and pixel analysis. Finally, all the features are digitalized to simple variables to represent their strength. Artificial Neural Network is used for the final facial expression classification.

The paper is organized as follows. Section 3 introduces basic concepts of ASM. Detailed digitalized facial feature extraction is described in Section 4. In Section 5 the comparisons and experiment results are discussed, and the conclusions are given in Section 6.

2. RELATED WORKS

Z. Yu suggested to use ASM as facial landmarks locating method, with RS-SVM for selection and classification [12]. RS-SVM reduced the computational complexity. However, its attribute reduction stage was not specifically designed for facial features. Detailed local facial feature extraction methods are not discussed.

R. Shibib developed a whole set of methods for facial expression recognition [13]. The facial features were extracted directly from ASM, and were not processed. An example in Chapter 3 shows only using ASM may cause confusions.

C. Hsieh and M. Jiang focused on local facial features, some regions of interest (ROI) were discussed [14]. However, the final extracted data are still in pixel format which increases the computing load of the following classifier.

All the above research agree that ASM is essential for FER. More detailed locations of ROI and more simplified extracted data are needed for FER.

3. BASIC CONCEPT ABOUT ACTIVE SHAPE MODEL

Active Shape Model (ASM) detects the profile of an object, it can be used to detect the shape of face and the parts of the face. To perform ASM to the face, a set of training samples of faces need to be labelled with landmark positions. Then apply Principal Component Analysis (PCA) to the data to find the eigenvalue and eigenvectors. Using different eigenvectors to form the new shapes which are limited by the eigenvalues.

The main idea of using ASM here is to locate the facial landmarks, so the locations of the other useful features can be found. With edge detection and pixel analysis, those useful features are accurately and easily extracted. Features such as the shapes of the nose and eyes will not be extracted because they are less related to facial expression, and this makes sure only the useful features are extracted.

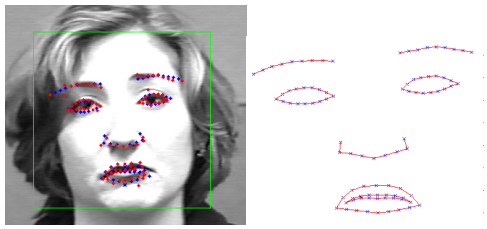

Figure 1. ASM Landmarks Locating

4. DIGITALISED FACIAL FEATURE EXTRACTION

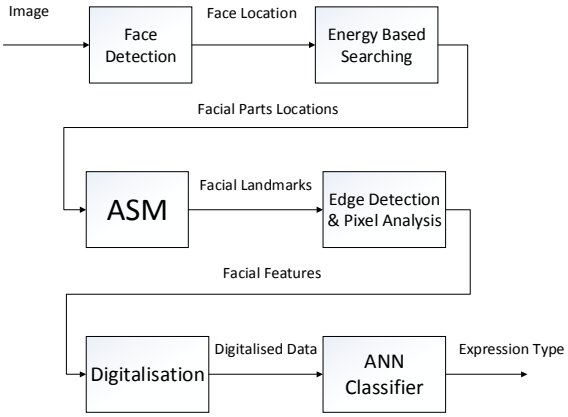

Figure 2. DFFA Flow Chart

The flow chart shows the whole progress of facial expression recognition using digitalized facial feature extraction. Face detection will be performed when an image is coming in, the rough face location will be found. Energy based searching will find more detailed facial parts locations. Then ASM will be applied to the facial image to search for the facial landmark. Edge detection and pixel analysis will extract the wrinkle features and eyeball feature. Then the useful features are digitalized. Passing through ANN classifier, the facial expression is recognized.

4.1. Energy Based Searching

Pre-processing of the facial image is very necessary for landmarks locating. In Face Detection stage, Haar-like features are used for face detection, however, only knowing the rough location of the face is not enough for ASM searching, more details need to be given to make sure the searching will not be trapped into local optima. So Energy Based Searching is used here to find out the rough centers of the key facial parts.

The facial image will first pass to a Gaussian filter to make the image smooth. Then accumulate the pixels horizontally and vertically. At last, calculate the first derivative and second derivative to find the centers. Figure 3 shows the original facial image and the plot of first derivatives in horizontal and vertical for Energy Based Searching.

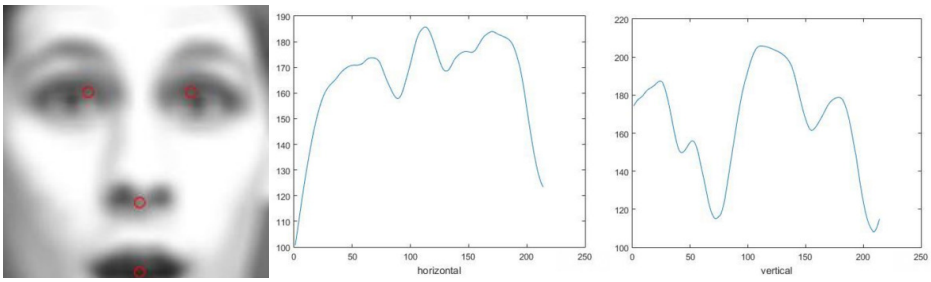

Figure 3. Energy Based Searching for Facial Parts Centers

By knowing the rough locations of the centers of eyes, nose and mouth, ASM searching is further limited. More accurate ASM results are assured.

4.2. ASM landmarks locating

When the face is detected, and the centers of facial parts are located, ASM is applied to the facial image. By using ASM, the profile of the facial parts are found. However, the ASM result cannot be used directly for expression recognition, here is an example.

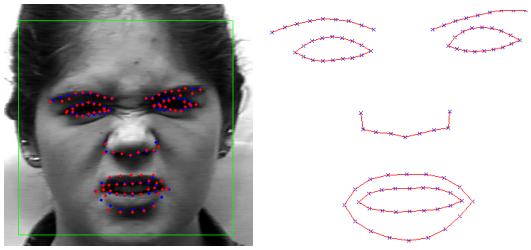

Figure 4. Confusing Expression with Only ASM Shapes without Wrinkles

Figure 4 shows a disgust expression, but if without wrinkles and using the ASM result only, it is difficult to tell the exact expression. So only using ASM is not capable of recognizing an expression. That means apart from ASM result, there is some other important information which are not extracted.

4.3. Edge Detection and Pixel Analysis

From Figure 4, it can be seen not the whole face is sending expression message. Dealing with parts which send no information is a waste. By using ASM for locating, the useful locations can be found. They are the locations of the forehead, between eyes and beside the nose. Wrinkles in those locations are crucial for FER. Thanks to ASM landmarks, these locations are already found. Edge Detection with Sobel Operator can easily detect the wrinkles.

Another useful information for expression is the eyeball black/white radio and the coverage of the upper and lower eyeball. ASM had already located the eye area, so the work here is to calculate the pixel inside the eye for analysis.

4.4. Digitalization of the Features

Wrinkle features for different people vary, but sometimes different wrinkles are giving the same expression. Simply using pixels or LBP or Gabor filters for feature extraction will have very different results. If these wrinkle features can be just described in a variable with a scalar to show the strength of the wrinkle, the problem is solved. It is the further digitalization of the features.
After the landmarks are found by ASM, five areas are the regions of interest (ROI). The forehead for wrinkle features (Area1), between eyes for wrinkle features (Area2), the eye feature (Area3), the mouth feature (Area4) and beside nose wrinkle features (Area5).

Features in the above 5 areas will be extracted and digitalized. In different areas, the features are selected differently. As the image may be stretched, the distance between centers of eyes and the distance between the end of nose to the point in the middle of eyes are used as a reference for horizontal and vertical.

To digitalize the wrinkles, the area needs to be defined. The digitalized wrinkle Dw is expressed as

$$Dw = \frac{\sum P_{an}}{N_P} \tag{1}$$

Where P_{an} is the magnitude of pixels which are above the threshold of Edge Detection, N_P is the total number of pixels in the area.

To digitalize the eye feature, a global average grey scale factor G_R for the face area need to be calculated. Eyeball black/white ratio De_r is expressed as

$$De_r = \frac{N_E - N_{Pw}}{N_{Pw}} \tag{2}$$

Where N_E is the total number of pixels in the eyes area, N_{Pw} is the number of pixels which are lighter than the average factor G_R.

The eyeball upper coverage De_u feature is expressed as

$$De_u = \frac{N_{Pub}}{N_{Eh}} \tag{3}$$

Where N_{Pub} is the number of pixels of the upper eyeball which are black. N_{Eh} is the number of pixels of half eyeball.

The eyeball lower coverage De_l feature is expressed as

$$De_l = \frac{N_{Plb}}{N_{Eh}} \tag{4}$$

Where N_{Plb} is the number of pixels of upper eyeball which are black.

Considering all the faces are highly symmetrical, only the right eye feature is used for digitalization.

To digitalize the mouth features D_m , four curvatures are used, the value of curvatures are from ASM result.

Finally, the facial features are expressed by

$$F = [\mathrm{Dw}_{\mathrm{f}}, \mathrm{Dw}_{\mathrm{be}}, \mathrm{Dw}_{\mathrm{bn}}, \mathrm{De}_{\mathrm{r}}, \mathrm{De}_{\mathrm{u}}, \mathrm{De}_{\mathrm{l}}, \mathrm{Dm}_{1}, \mathrm{Dm}_{2}, \mathrm{Dm}_{3}, \mathrm{Dm}_{4}] \quad (5)$$

5. EXPERIMENTS AND RESULTS

5.1. Experiment Conditions and Training

The facial expression database used for training is Cohn-Kanade (CK) database [15, 16]. Compared with other facial expression databases like Japanese Female Facial Expression (JAFFE) database which only has 213 faces from 7 women, CK database has better diversity and more images.

Figure 5: Example of Cohn-Kanade database

150 of images are chosen randomly from the CK database with six basic expressions and neutral expression equally, 101 of them are used for training, and 49 of them are for verification.

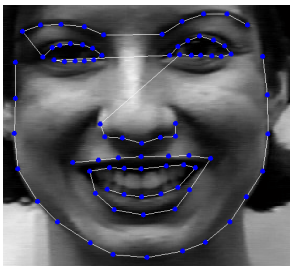

Figure 6. 82 points ASM landmarks

5.2. Comparison with Other Feature Extraction Methods

For DFFA, five key areas with ten parameters are extracted, the ASM searching uses 82 key landmarks. For Gabor filters, eight directions and five scales are used.

The first comparison is about the computational complexity of the extracted features for the following classifier between different feature extraction methods and different resolutions.

Table 1. Parameters after Extraction

Resolution	DFFA	Gabor Filters (8d5s)	LBP
64x64	10	1638400	4096
128x128	10	6553600	16384

Table 1 shows, after extraction, the computational complexity of DFFA is much lower, such scale of Gabor wavelet features are almost unable to be processed in real-time.

The next comparison is the recognition rate between the different feature extraction methods, the classifier is ANN.

Table 2. Feature Extraction Methods Comparison

Type of Method	Recognition Rate
Gabor Features	89.87%
LBP	71.4%
DFFA	85.7%

The result shows DFFA has acceptable recognition rate.

6. CONCLUSION

DFFA shows high performance in FER and is suitable for real-time FER. It reduces the computational complexity while keeping acceptable recognition rate. Thanks to the digitalization of facial features and edge detections, the huge facial feature data were extracted and only the most useful feature information is converted into small scaled data set. Future works will focus on a high-efficiency classifier design for the digitalized data.

REFERENCES

[1] P. Ekman, W. Friesen, Facial action coding system, Palo Alto CA, USA, Consulting Psychologist Press, 1978.

[2] T. Ojala, M Pietikinen, and D. Harwood, “A comparative study of texture measures with classification based on featured distribution,” Pattern Recognition, vol. 29, no. 1, 1996.

[3] T. Ojala, M. Pietikinen, and T. Menp, “Multiresolution grayscale and rotation invariant texture classification with local binary patterns,” IEEE PAMI, vol. 24, no. 7, July 2002.

[4] T. Ahonen, A. Hadid, and M. Pietikinen, “Face recognition with local binary patterns,” ECCV, 2004, pp. 469–481.

[5] A. Hadid, M. Pietikinen, and T. Ahonen, “A discriminative feature space for detecting and recognizing faces,” IEEE CVPR, June 2004, pp. 797–804.

[6] X. Tan, B. Triggs, “Enhanced Local Texture Feature Sets for Face Recognition Under Difficult Lighting Conditions”, IEEE Trans. Image Processing, Vol. 19, No. 6, pp. 1635-1650, June 2010.

[7] M. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan, “Recognizing facial expression: machine learning and application to spontaneous behavior,” Computer Vision and Pattern Recognition, pp. 568-573, vol. 2, 2005.

[8] L. Ma, D. Chelberg, and M. Celenk. "Spatio-temporal modeling of facial expressions using Gabor-wavelets and hierarchical hidden Markov models." IEEE International Conference on Image Processing 2005. Vol. 2, 2005.

[9] T. F. Cootes and C. J.Taylor. “Active shape models”. 3rd British Machine Vision Conference 1992, pages 266–275, 1992.

[10] B. van Ginneken, A.F. Frangi, J.J. Staal, B.M. ter Har Romeny, and M.A. Viergever, “Active shape model segmentation with optimal features,” IEEE Trans. Medical Imaging, vol. 21, no. 8, pp. 924-933, 2002.

[11] F.M. Sunko, S. Ordaas, C. Butakoff, S. Cruz, and A.F. Frangi, “Active shape models with invariant optimal features: Application to Facial Analysis,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 29, no. 7, pp. 1105-1117, 2007.

[12] Yu Zheng-Hong, Li Cong. “Research of Facial Expression Recognition Based on ASM Model and RS-SVM”. ISDEA, pp. 772-777, 2014.

[13] Shbib, Reda, and Shikun Zhou. "Facial expression analysis using active shape model." Int. J. Signal Process. Image Process. Pattern Recognition, vol. 8, no. 1, pp-9-22, 2015.

[14] Hsieh, Chen-Chiung, and Meng-Kai Jiang. "A facial expression classification system based on active shape model and support vector machine." Computer Science and Society (ISCCS), 2011 International Symposium on. IEEE, pp. 311-314, 2011.

[15] Kanade, T., Cohn, J. F., & Tian, Y. (2000). Comprehensive database for facial expression analysis. Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition (FG'00), Grenoble, France, 46-53.

[16] Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I. (2010). The Extended Cohn-Kanade Dataset (CK+): A complete expression dataset for action unit and emotion-specified expression. Proceedings of the Third International Workshop on CVPR for Human Communicative Behavior Analysis (CVPR4HB 2010), San Francisco, USA, 94-101.

DIGITAL VIDEO SOURCE IDENTIFICATION BASED ON GREEN-CHANNEL PHOTO RESPONSE NON-UNIFORMITY (G-PRNU)

M.Al-Athamneh[1], F.Kurugollu[2], D.Crookes[3] and M. Farid[4]

The Institute of Electronics, Communications and Information Technology (ECIT),
Queen's University Belfast.
[123]{malathamneh01, f.kurugollu, d.crookes} @qub.ac.uk
[4]Department of Electronics, Computing and Mathematics
University of Derby
m.farid@derby.ac.uk

ABSTRACT

This paper proposes a simple but yet an effective new method for the problem of digital video camera identification. It is known that after an exposure time of 0.15 seconds, the green channel is the noisiest of the three RGB colour channels [5]. Based on this observation, the digital camera pattern noise reference, which is extracted using only the green channel of the frames and is called Green-channel Photo Response Non-Uniformity (G-PRNU), is exploited as a fingerprint of the camera. The green channels are first resized to a standard frame size (512x512) using bilinear interpolation. Then the camera fingerprint is obtained by a wavelet based denoising filter described in [4] and averaged over the frames. 2-D correlation coefficient is used in the detection test. This method has been evaluated using 290 video sequences taken by four consumer digital video cameras and two mobile phones. The results show G-PRNU has potential to be a reliable technique in digital video camera identification, and gives better results than PRNU.

KEYWORDS

PRNU, G-PRNU, Video Forensics.

1. INTRODUCTION

The first legislation which recognized computer crime was in 1978 (Florida Computer Crimes Act), which was against the unauthorized modification or deletion of data on a computer system. Since then, 'Digital Evidence' has become a new type of evidence in the judicial system.

Digital evidence has increased tremendously in the last few decades, as courts of law allow the use of e-mails, digital photographs, digital video or audio files, ATM transaction logs, word processing documents, spread-sheets, internet browser histories, computer memory contents, computer backups, and Global Positioning Systems tracks.

Dhinaharan Nagamalai et al. (Eds) : ICCSEA, SPPR, UBIC - 2016
pp. 47–57, 2016. © CS & IT-CSCP 2016 DOI : 10.5121/csit.2016.61105

As with any other type of evidence presented to the court of law, digital evidence is subjected to integrity and authenticity checks. An integrity check aims to ensure that the act of seizing did not modify the evidence; authenticity refers to the ability to confirm the trustworthiness of presented evidence; e.g. to show it has not been tampered with.

Digital videos were introduced commercially in 1986 with the Sony D-1 format. Since then they have experienced enormous growth, and have been used in a growing number of applications. They can be found everywhere in today's daily life like consumer digital video cameras, mobile phones, CCTV cameras, DVDs, internet, etc. Research topics in digital video forensics include source camera identification, device linking, authentication, integrity verification, etc. It has become an emerging field due to the availability of sophisticated video editing tools, and because of the lack of methodologies for validating the source of digital videos.

Identifying the device used in acquiring a particular digital video is important, as it can be used as a definitive proof (or disproof) of events in a court of law. It can be likened to firearm forensics, in which the imperfections in the surface of the interior of the barrel create cracks on the projectiles which produce a unique "bullet scratch" pattern on every bullet that passes through the barrel of the firearm. The equivalent of "bullet scratch" in digital forensics is the Photo Response Non-Uniformity (PRNU).

Extracting PRNU is very sensitive process as the PRNU is a weak signal and can be manipulated easily by the content in the digital data, previous research has shown that by including a weighting factor during PRNU extraction can improve the correct identification, by using the weighting factor smooth regions are emphasized, edges and highly-texture regions are deemphasized during the denoising process [23].

Other research worked on enhancing the estimation of PRNU using probabilistically estimated raw data to obtain better camera identification for small dimension patches, Poisson process and Maximum Likelihood Estimation (MLE) is used to extract the PRNU [21].

Not only in the field of source identification and manipulation detection , PRNU can also be used to improve a biometric systems Security by ensuring the authenticity and integrity of images acquired with a biometric sensor [22].

In a digital video camera the light passes through a colour filter array (CFA), which is positioned over the sensor to separate out the red, green, and blue components of light falling on it. The GRGB Bayer Pattern is the most common CFA used in digital video cameras as depicted in Figure 1; this process produces the digital video frame, which can be represented by a matrix of RGB values.

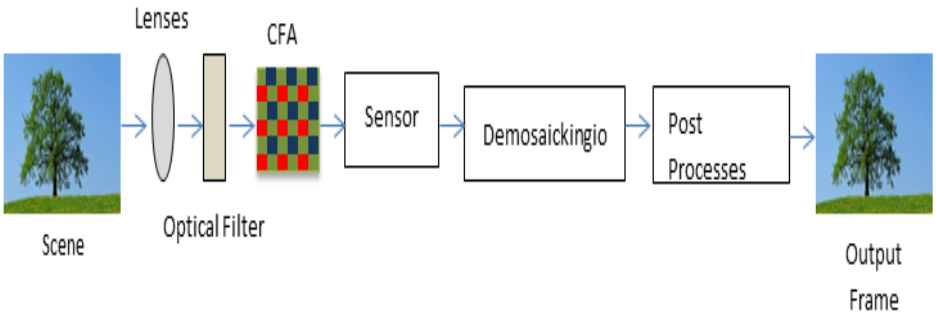

Figure 1. The digital video frame acquisition process.

Previous research has shown that the relationship between the PRNU and the exposure time is approximately linear, and that the green channel has the highest PRNU, followed by the red and then the blue channels respectively. After an exposure time of 0.15s the green channel is the noisiest channel among the three colours RGB [5].

In this paper, a new method for digital video source identification using Green-Channel PRNU (G-PRNU) is proposed. First the frames are resized to 512x512 using bilinear interpolation, to facilitate calculating 2-D correlation between different video sequences. Then the noise in the resized frames is obtained using a wavelet-based denoising filter [4]. The camera fingerprint is obtained by averaging the noise over all frames. Performance of green channel only PRNU in video sequences is superior to ordinary PRNU. Moreover the use of bilinear interpolation for resizing also improves the performance of the proposed method.

The rest of the paper is organised as follows: PRNU is introduced in the next section. Section 3 presents G-PRNU and explains how the camera reference is extracted. The bilinear interpolation and noise detection strategy are also discussed in this section. The results are presented in Section 4 and the conclusions are drawn in Section 5.

2. PHOTO RESPONSE NON-UNIFORMITY (PRNU)

PRNU can be simply defined as: The imperfections of manufacturing semiconductor wafers, and the variations in which individual sensor pixels create a sensor-specific noise pattern which makes it possible to identify the imaging source. Its reliability in identifying devices and related models has been demonstrated [3,9,10,12,14,18,19,20].

The PRNU method creates the digital camera reference (the consistent noise pattern of each sensor pixel) by finding the noise in captured images, and averaging this noise over a set of images to give the camera reference. The noise can be estimated by denoising the original images.

PRNU has proved to be an effective technique for determining the source of a digital image (the ‘suspected image’) using the following steps:

- Create the digital camera reference, by averaging the noise obtained by, for example, a Wavelet based denoising algorithm.

- Create the suspected image reference: this image reference is the noise obtained by using the previous denoising technique.

- Finally, apply a camera reference detection method, by matching the suspected image reference against a set of camera reference images. The main detection methods are: normalized correlation coefficient, the peak to correlation energy (PCE) and correlation to circular correlation norm (CCN) [9,16].

3. G-PRNU CAMERA REFERENCE

3.1. G-PRNU EXTRACTION

Most previous research has focused on still images, while modern forensic applications are mostly based on digital videos rather than still images (CCTV, camcorders, cell-phones, etc.); therefore our aim in this paper is to create an effective method to determine the source of a digital video. The motivation behind this is the need to authenticate digital video evidence and prove its trustworthiness in a court of law.

Because the green colour channel of video frames is normally noisier than the other colour channels, G-PRNU (Green - Photo Response Non-Uniformity) was created by examining the green colour channel of the video frames.

$$\boldsymbol{A'_k = f(A_k)} \qquad (1)$$

Where A'_k is the denoised digital video frame (k),$A_k$is the actual captured frame and f is the denoising function [4].

The actual noise can then be extracted from frame k:

$$\boldsymbol{N_k = A_k - A'_k} \qquad (2)$$

Where N_k is the frame (k) noise. Then we average this over a set of captured frames:

$$\boldsymbol{y = \frac{\sum_{k=1}^{l} N_k}{l}} \qquad (3)$$

Where y is the digital video G-PRNU reference, l is the total number of frames (in our case, 350 frames per video).

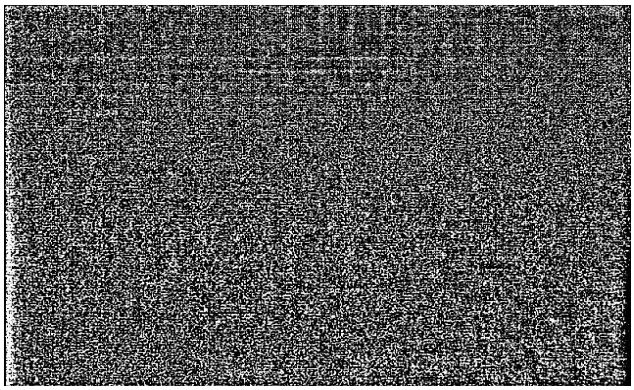

Figure 2. G-PRNU reference of Canon IXUS 8515 (C2), Note: The reference image is x6 scaled for viewing purpose.

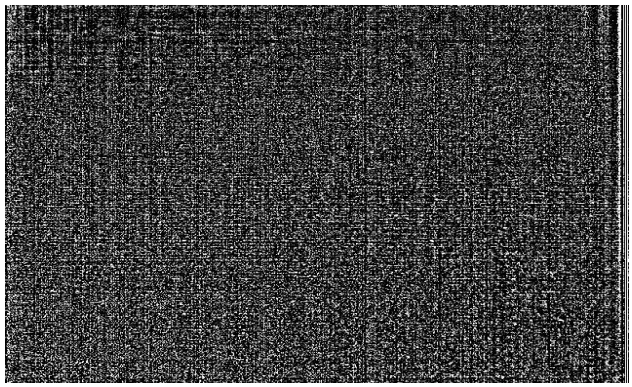

Figure 3. G-PRNU reference of Huawei Ascends G300 (C6), Note: reference image is x6 scaled for viewing purpose.

The difference in G-PRNU camera reference for two different cameras can be visually noted: in Figure 2, note the greater level of noise at the left edge of the reference image, while in Figure 3 the noise is more smoothly distributed with little bit more noise on the right edge.

3.2. FRAME RESIZING

When comparing videos potentially from a range of different cameras, the question of frame size was encountered. Even a single camera can produce a range of frame sizes. Also, the suspected videos may have been resized. Therefore, all reference images are resized to a standard size (for example, 512x512 is used), for comparison using 2-D correlation. This should be carried out with extra care avoiding any damage of the characteristic noise patterns.

Bilinear Interpolation which is an extension of linear interpolation was used in this work where the output (interpolated) pixel value is a weighted average of pixels in the nearest 2x2 neighbourhood, This results in much smoother looking images than Nearest Neighbor and Bicubic interpolation which they were experimentally used in this work but yet Bilinear Interpolation proved to give the best correlation results [Table 3].

CALCULATING BILINEAR INTERPOLATION

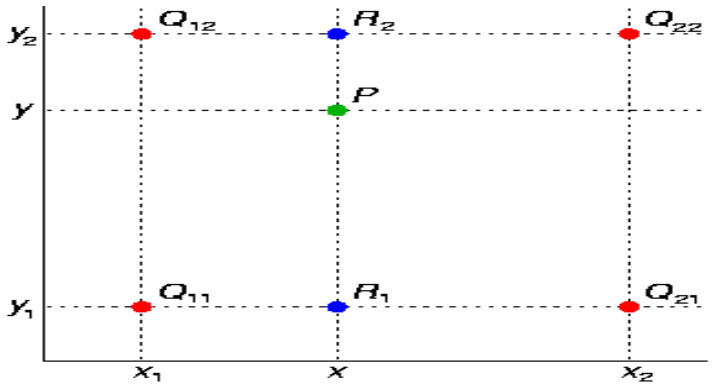

Figure 4. Bilinear Interpolation calculating method

There are several equivalent ways to calculate the value of the bilinear interpolation P. An easy way to calculate the value of P would be to first calculate the value of the two blue dots, R2, and R1. R2 is effectively a weighted average of Q12 and Q22, while R1 is a weighted average of Q11 and Q21.

$$R1 = ((x2 - x)/(x2 - x1))*Q11 + ((x - x1)/(x2 - x1))*Q21$$

$$R2 = ((x2 - x)/(x2 - x1))*Q12 + ((x - x1)/(x2 - x1))*Q22$$

After the two R values are calculated, the value of P can finally be calculated by a weighted average of R1 and R2.

$$P = ((y2 - y)/(y2 - y1))*R1 + ((y - y1)/(y2 - y1))*R2$$

There are two main reasons for using bilinear interpolation over other methods: first it helps to get better correlation [Table 2], and second it solves the matrix dimension mismatch problem which deters the calculation of 2-D correlation coefficients.

3.3. G-PRNU DETECTION

Applying an effective detection method for camera identification is no less important than constructing the camera reference in the first place. Some researchers have used the normalized correlation coefficient for this purpose [1,2,6] while others have shown that the peak to correlation energy (PCE) can give better results in detection tests [8,9]. Furthermore circular correlation norm (CCN) [10] has also been used as a statistical detection test which basically enhances the results of PCE by lowering the false positive rates.

In this research the 2-D correlation coefficient was used in the camera detection test, which computes the correlation coefficient between the video camera reference (A) and the suspected video reference (B), where A and B are G-PRNU matrices of the same size.

$$\mathbf{r} = \frac{\sum_m \sum_n (\mathbf{A_{mn}} - \overline{\mathbf{A}})(\mathbf{B_{mn}} - \overline{\mathbf{B}})}{\sqrt{\left(\sum_m \sum_n (\mathbf{A_{mn}} - \overline{\mathbf{A}})^2\right)\left(\sum_m \sum_n (\mathbf{B_{mn}} - \overline{\mathbf{B}})^2\right)}} \qquad (4)$$

$\overline{A}$ and $\overline{B}$ are the mean of A, B respectively.

Overall the proposed G-PRNU method can be summarised as follows:

1. Extract the green channel frames from the video (350 frames/video).
2. Resize the extracted frames to [512x512] using bilinear interpolation.
3. Perform wavelet-based de-noising on these green channel frames (denoising by soft-thresholding [4]).
4. Create the G-PRNU map for the video by averaging the results of step 3.
5. To create a camera reference, perform steps1-4 on 9 videos captured by the same camera.
6. Use the 2-D correlation coefficient as the camera detection test.

4. EXPERIMENTAL RESULTS

First, a dataset of videos taken by a range of consumer video cameras was constructed. To try to simulate the real life cases of digital video forensics, four video cameras and two mobile phones (Samsung GT-S5830 and Huawei Ascend G300) [Table 1] were considered. Over a period of four months, these cameras were used to capture a total of 254 videos under various conditions: indoors, outdoors, sunny days, rainy days, good lighting and poor lighting; some videos were captured by two different cameras at the same time by holding them side by side.

The camera reference for each of the six cameras was calculated by selecting nine videos from each camera; the remaining 236 videos used as the test data set.

The 2-D correlation coefficient detection test was applied to identify the source of each of the 236 test videos, by matching against all six video references, using PRNU, G-PRNU, and G-PRNU interpolated by 512x512 bilinear interpolations. The results are shown in Table 2. This shows the cameras' identification rates and proves that the G-PRNU approach gives considerable success in identifying the source device of digital video. From a total of 236 test videos, G-PRNU could correctly determine the source of 234 videos (correct detection rate was 99.15%);in comparison, using PCE [9], the correct detection rate was 41.15%.Note, in creating PRNU references, the same method was used as in creating the G-PRNU references.

Figure 5 shows the 2-D correlation of videos captured by C4 versus the six cameras G-PRNU fingerprints, all videos in figure 5 have obtained higher correlation with the C4 G-PRNU fingerprint, this led to the conclusion that videos in figure 5 were captured by C4,figures 6 and 7 describe the same situation for C5 and C6, respectively.

Table 1: Cameras used in G-PRNU experiments

Symbol	Camera Brand	Sensor	Format
C1	Fujifilm F550EXR	1/2" (6.4 x 4.8 mm) EXRCMOS	.mov
C2	Canon IXUS 8515	1/2.3" (6.17 x 4.55 mm) CCD	.avi
C3	Samsung GT-S5830	Unknown	.mp4
C4	Canon SD1000	1/2.5" (5.744 x 4.308 mm) CCD	.avi
C5	Fujifilm jv2000	1/2.3" (6.17 x 4.55 mm) CCD	.avi
C6	Huawei Ascend G300	Unknown	.mp4

Table 2: Source camera identification rates

Camera	PRNU	G-PRNU	G-PRNU with Bilinear interpolation
C1	15%	97.5%	**100%**
C2	36.58%	95.12%	**97.56%**
C3	25.8%	96.77%	**97.56%**
C4	37.83%	100%	**100%**
C5	26.47%	100%	**100%**
C6	95.34%	97.67%	**100%**
Total	41.15%	97.79%	**99.15%**

Table 3: Source camera successful identification rates using deferent Interpolations and Dimensions

Interpolation	Dimension				
	64x64	128x128	256x256	512x512	640x640
Bicubic	76.51	82.53	88.55	92.77	92.17
Bilinear	72.29	82.53	87.35	**99.15**	93.37
Nearest	71.69	80.72	85.96	88.55	79.52

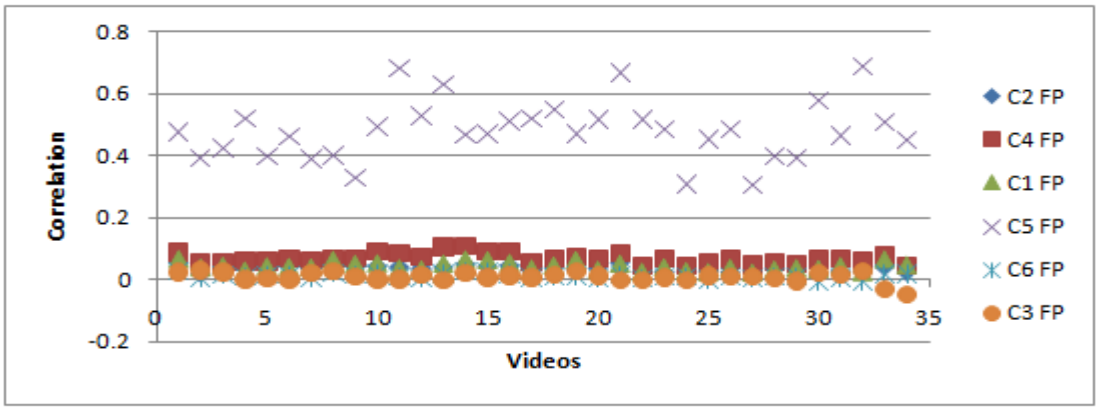

Figure 5. G-PRNU Correlation of videos captured by C5.

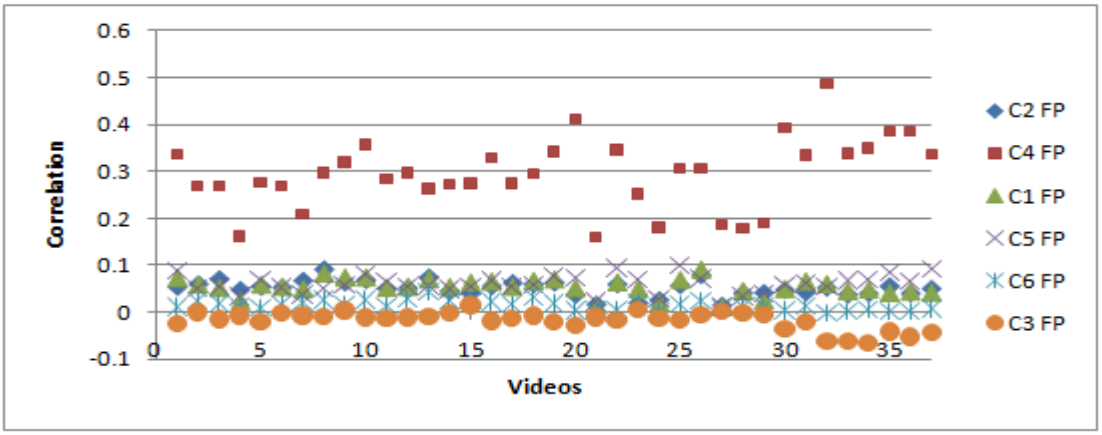

Figure 6. G-PRNU Correlation of videos captured by C4.

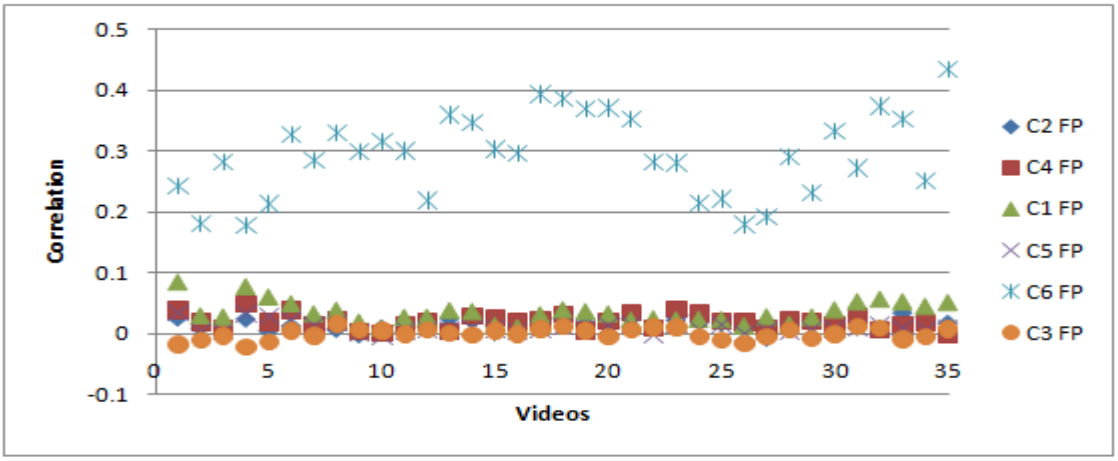

Figure 7. G-PRNU Correlation of videos captured by C6.

5. CONCLUSIONS

A new method for video camera identification has been proposed, called G-PRNU. The method was tested using six cameras (4 consumer cameras and two mobile phones). 290 videos were captured under various conditions. In initial experiments, the G-PRNU method showed potential to be a reliable technique in digital video camera identification and proven to give better results than PRNU in the problem of digital videos source identification.

Also the question of frame size was encountered, as even a single camera can produce a range of frame sizes, Frame resizing should be carried out with extra care avoiding any damage of the characteristic noise patterns. The Bilinear Interpolation which is an extension of linear interpolation gave results much smoother looking images and correct detection results than Nearest Neighbor and Bicubic interpolation.

REFERENCES

[1] A. Popescu, and H. Farid, “Exposing Digital Forgeries in Color Filter Array Interpolated Images", IEEE Transactions on Signal Processing, vol. 53, no. 10, pp.3948-3959,2005.

[2] A. Swaminathan, Min Wu, Liu, K.J.R., “Digital Image Forensics via Intrinsic Fingerprints”, IEEE transactions on information forensics and security, vol. 3, Iss. 1, pp. 101-117, 2008.

[3] C. Li, "Source Camera Identification Using Enhanced Sensor Pattern Noise," IEEE Transactions on Information Forensics and Security, vol. 5, no. 2, June 2010.

[4] D.L. Donoho, "De-noising by soft-thresholding," IEEE transactions on information theory, vol. 41, no. 3, pp. 613–627, MAY 1995

[5] http://scien.stanford.edu/pages/labsite/2005/psych221/projects/05/joanmoh/prnu.html, [accessed Jan 2014]

[6] http://dde.binghamton.edu/download/, [accessed Jan 2014]

[7] J. R. Janesick,“Scientific Charge-Coupled Devices”, Bellingham, WA: SPIE, vol. PM83, 2001.

[8] J. Lukas, J. Fridrich, and M. Goljan, “Digital "bullet scratches" for images“, International Conference on Image Processing (ICIP), IEEE, vol. 31, pp.65 – 68,2005.

[9] J. Lukas, J. Fridrich, and M. Goljan, "Digital Camera Identification from Sensor Noise," IEEE Transactions on Information Security and Forensics, vol. 1, no. 2, pp. 205-214, 2006.

[10] J. Fridrich, “Digital image forensics”, IEEE signal processing, vol. 26, Iss. 2, pp.26-37, 2009.

[11] K.S Thyagarajan “Still image and video compression with MATLAB”, Wiley, 1st ed., 428p,2010.

[12] M. Chen, J. Fridrich, M. Goljan, and J. Lukas, "Determining Image Origin and Integrity Using Sensor Noise," IEEE Transactions on Information Security and Forensics, vol. 3, no. 1, pp. 74-90, 2008.

[13] M. Goljan, J. Fridrich, and T. Filler, “Large scale test of sensor fingerprint camera identification,” in Proc. SPIE, San Jose, CA, Jan. 18–22,2009, vol. 7254, pp. 0I 1–0I 12, Electronic Imaging, Media Forensics and Security XI.

[14] M. Goljan and J. Fridrich, “Digital camera identification from images Estimating false acceptance probability” in Proc. 8th Int.Workshop Digital Watermarking, Busan, Korea, Nov. 10–12, 2008.

[15] R.C. Gonzalez, R.Woods, S. Eddins, “Digital image processing using MATLAB”, Gatesmark Publishing, 2nd ed., 827p, 2009.

[16] X. Kang, Y. Li, Z. Qu, and J. Huang, “Enhancing Source Camera Identification Performance with a Camera Reference Phase Sensor Pattern Noise”, IEEE Trans. Inform. Forensics Security, vol. 1, APRIL 2012.

[17] Y. Q.Shi, and H. Sun, “Image and video compression for multimedia engineering: fundamentals, algorithms, and standards”, CRC Press LLC, 480p, 2000.

[18] Min-Jen Tsai, Cheng-Liang Lai, Jung Liu, “Camera/Mobile Phone Source Identification for Digital Forensics”, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Apr. 2007, Hawaii USA.

[19] E. Dirik, H. T. Sencar, and N. Memon, “Source camera identification based on sensor dust characteristics,” in Proc. Signal Processing Applications Public Security Forensics, Apr. 11–13,2007, pp. 1–6.

[20] Y. Long and Y. Huang, “Image based source camera identification using demosaicking,” in Proc. IEEE 8th Workshop Multimedia Signal Processing, Oct. 2006, pp. 419–424.

[21] A. Mehrish, A. V. Subramanyam and S. Emmanuel, "Sensor Pattern Noise Estimation Using Probabilistically Estimated RAW Values," in IEEE Signal Processing Letters, vol. 23, no. 5, pp. 693-697, May 2016.

[22] L. Debiasi and A. Uhl, "Comparison of PRNU enhancement techniques to generate PRNU fingerprints for biometric source sensor attribution," 2016 4th International Conference on Biometrics and Forensics (IWBF), Limassol, 2016, pp. 1-6.

[23] K. f. Lau, N. f. Law and W. c. Siu, "Use of sensor noise for source identification," Noise and Fluctuations (ICNF), 2015 International Conference on, Xian, 2015, pp. 1-6.

INTENTIONAL BLANK

WIRELESS SENSORS INTEGRATION INTO INTERNET OF THINGS AND THE SECURITY PRIMITIVES

Muhammad A. Iqbal and Dr. Magdy Bayoumi

The Center for Advanced Computer Studies, University of Louisiana at Lafayette, Lafayette, LA 70504 USA
mxi1678@cacs.louisiana.edu, mab@cacs.louisiana.edu

ABSTRACT

The common vision of smart systems today, is by and large associated with one single concept, the internet of things (IoT), where the whole physical infrastructure is linked with intelligent monitoring and communication technologies through the use of wireless sensors. In such an intelligent vibrant system, sensors are connected to send useful information and control instructions via distributed sensor networks. Wireless sensors have an easy deployment and better flexibility of devices contrary to wired setup. With the rapid technological development of sensors, wireless sensor networks (WSNs) will become the key technology for IoT and an invaluable resource for realizing the vision of Internet of things (IoT) paradigm. It is also important to consider whether the sensors of a WSN should be completely integrated into IoT or not. New security challenges arise when heterogeneous sensors are integrated into the IoT. Security needs to be considered at a global perspective, not just at a local scale. This paper gives an overview of sensor integration into IoT, some major security challenges and also a number of security primitives that can be taken to protect their data over the internet.

KEYWORDS

Internet of Things (IoT), Wireless Sensor Network (WSN), Security, Privacy, Integration, Confidentiality

1. INTRODUCTION AND RELATED WORK

Today, Internet of things (IoT) itself has become a thing – a thing worth talking about, from the university project discussions to conferences to giant tech companies' meetings. IoT is being identified as one of the top emerging future technologies. The concept is simple at its core; connecting devices over the internet: making them 'smart'. We can think of it as the internet expanding from being a network of computers to a network of both computers and things. This idea is not even new, indeed first 'thing' connected to internet was a Coke vending machine by Carnegie Mellon University students in 1982. What is new added into this concept, are the sensors - tiny sensors embedded in devices that can gather almost any kind of information about their surrounding environment (temperature, light, sound, time, movement, speed, distance, and more).

The term internet of things was devised by Kevin Ashton in 1999, co-founder and executive director of Auto-ID Center at MIT and refers to uniquely identifiable objects and their virtual

Dhinaharan Nagamalai et al. (Eds) : ICCSEA, SPPR, UBIC - 2016
pp. 59–67, 2016. © CS & IT-CSCP 2016 DOI : 10.5121/csit.2016.61106

representations in an "internet-like" structure. With the advancement in technology, the cost of sensors, processors and transmitters is becoming less and their computational and processing powers becoming higher, allowing putting them into any object of our day-to-day life i.e. the food, the clothes, the medicines and so on. The technological advances also enhance this connectivity by adding one more dimension to it - connecting anything. Just to give an example, Nike has recently introduced a new line of running shoes that can track its wearer's progress and post updates online. The age of so called 'smart dust', which we have been talking about for years now, is finally upon us after the development of a fully functional computer with built-in wireless connectivity measuring just one cubic mm [2]. There is even a tree in Brussels, Belgium packed with sensors and cameras that constantly posts local environmental updates on Twitter. And that tree has 3,000 followers, how many people on Twitter can say they have 3,000 followers? At least I can't [2].

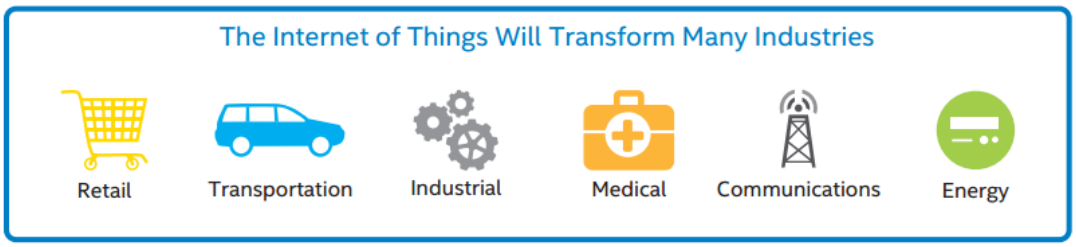

Figure 1. Internet of things Applications [14]

It is estimated that the number of connected devices is expected to grow exponentially to 50 billion by 2020 by Cisco. Intel, more optimistically, predicts 200 billion by that same year. The main driver for this growth is not human population; rather, the fact that devices we use in our daily life (e.g., refrigerators, cars, fans, lights) and operational technologies such as those found on the factory floor are becoming connected entities across the globe. This world of interconnected things - where the humans are interacting with the machines and machines are talking with other machines (M2M) — is here and it is here to stay [13].

An interesting trend contributing to the growth of IoT is shift from the consumer-based IPv4 Internet of tablets and laptops, that is, IT to Operational Technology (OT) based IPv6 Internet of M2M interactions. This includes sensors, smart objects and clustered systems (for example, Smart Grid). IPv6 is new enabling technology, an upgrade to the Internet's original fundamental protocol – the Internet Protocol (IP), which supports all communications on the Internet. IPv6 is necessary because the Internet is running out of original IPv4 addresses. Key challenge here is to make IPv6 interoperable for the most IoT software developed for IPv4 and readily available. Many experts believe, however, that IPv6 is the best connectivity option and will allow IoT to reach its potential.

Challenges that need to be addressed include how to communicate effectively and securely between devices, how to transmit and store huge amounts of data, and how to protect the privacy. A major barrier to realizing the full promise of IoT is that around 85% of existing things were not designed to connect to the Internet and cannot share data with the cloud [4]. Addressing this issue, gateways from mobile, home, and industry playing the part to act as intermediaries between legacy things and the cloud, not only providing the required connectivity but also security and the manageability [14] as shown in figure 2.

The things to be connected to the Internet largely vary in terms of characteristics. This ranges from very small and static devices (e.g., RFIDs) to large and mobile devices (e.g., vehicles). Such

heterogeneity induces complexity and stipulates the presence of an advanced middleware that can mask this heterogeneity and promote transparency. Among other technologies, radio

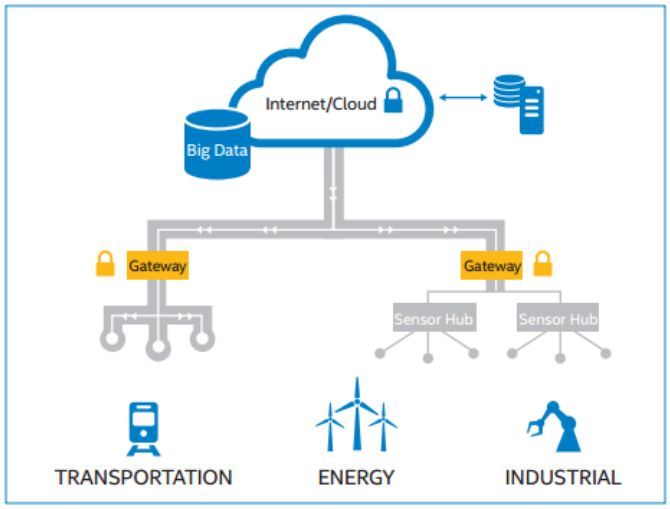

Figure 2. Addressing Endless Use Cases and Gateways [14]

frequency identification (RFID) and wireless sensor network (WSN) represent two of the most promising technologies enabling the implementation of IoT infrastructure. RFID is a low-cost, low-power technology consisting of passive or battery-assisted passive devices (tags) that are able to transmit data when powered by the electromagnetic field generated by an interrogator (reader). Since passive RFID tags do not need a source of energy to operate, their lifetime can be measured in decades, thus making the RFID technology well suited in a variety of application scenarios, including the industrial and healthcare ones [4]. The main challenges for RFID are non-uniform encoding, conflict collision and RFID privacy protection.

On the other hand, WSNs are basically self-organizing ad hoc networks of small, cost-effective devices (motes) that communicate/cooperate in a multi-hop fashion to provide monitor and control functionalities in critical applications including industrial, military, home, automotive, and healthcare scenarios. Currently, most WSN motes are battery-powered computing platforms integrating analogue/digital sensors and an IEEE 802.15.4 radio enabling up to 100m outdoor communication range (single hop). Unlike other networks, WSNs have the particular characteristic of collecting sensed data (temperature, motion, pressure, fire detection, voltage/current, etc) and forwarding it to the base station or gateway. Even though most WSN protocols were not designed for two-way communications as illustrated in IMS research, they should also be able to receive information and send it to the sensors (a command for example), and react on behalf of the commander/user, e.g., automating home appliances.

2. SENSOR NETWORKS IN A GLOBALLY CONNECTED NETWORK

Integration of Wireless Sensor Networks (WSN) into IoT is not mere speculation, a number of big technology companies supporting and developing their IoT infrastructure around WSN. Noteworthy examples are IBM's 'A Smarter Planet', a strategy considers sensors as fundamental pillars in intelligent water management systems and intelligent cities; and the CeNSE project by HP Labs, focused on the deployment of a worldwide sensor network in order to create a "central nervous system for the Earth" [11].

The question how sensor nodes should provide their services when connecting WSN to the Internet is important; either directly or through the base station. The 'thing' connected to Internet is required to be locatable and addressable via the Internet, but this particular configuration might not be suitable for certain scenarios. Some specific scenarios for instance, in SCADA systems a sensor node does not need to provide its services directly and other scenarios are where a sensor node should be completely integrated into the Internet.

There are different approaches in which WSNs are integrated into IoT depending upon the application requirements and already deployed WSNs infrastructure. Some of these approaches are described below:

- Completely independent from internet where WSN has its own protocol and sensor nodes communicate through a centralized device called base station. Any query coming from internet host is traversed through base station that collects and holds all data from sensor nodes. If base station acts as applications layer gateway, then Internet hosts and sensor nodes able to address each other and exchange information without establishing a true direct connection. WSN still independent of internet, all queries go through gateway.

- Sensor nodes implement TCP/IP stack (or a compatible set of protocols such as 6LoWPAN) so that any internet host can have direct communication with them & vice versa. Sensor nodes are no longer to use specific WSN protocols.

- There's another topology based integration approach in which level of integration depends on actual location of the nodes, nodes can be dual sensors (base stations) located on the root of the WSN or full-fledged backbone of devices that allow sensing nodes to access Internet in one-hop (access point). WSN becomes unbalanced tree with multiple roots, leaves are normal sensors nodes and other elements are internet-enabled nodes.

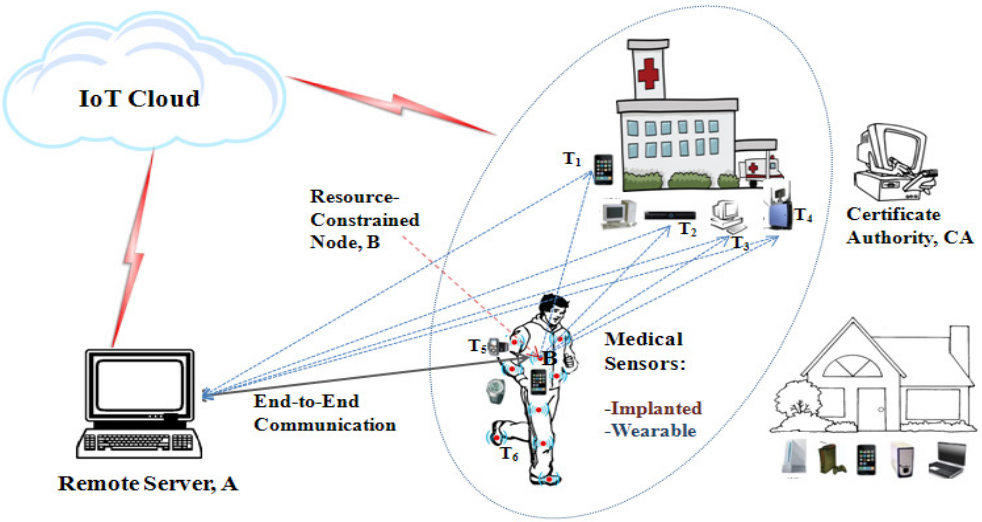

Figure 3. Typical Body Area Network Scenario in context of IoT

The evolution of the IoT has its origin in the convergence of wireless technologies, advancements of MEMS and digital electronics where as a result miniature devices with the ability to sense and compute are communicating wirelessly. However, having IP connectivity does not mean that every sensor node should be directly connected to the Internet. There are many challenges that must be carefully considered, and one of those challenges is security [11]. In the era of IoT, the

interaction or relationship between humans and machines needs to be considered more seriously as machines getting smarter and starting to handle more human tasks. A thing might be a patient with a medical implant to facilitate real-time monitoring in a healthcare application or an accelerometer for movement attached to the cow in a farm environment. In such a situation, humans are required to trust the machines and feel safe about it [10].

3. IOT SECURITY AND PRIVACY

IoT applications projections predict a safer, smarter and efficient world while some observers show concerns that it would be a darker world of surveillance, privacy and security violations, and consumer lock–in. The scale and context of the IoT make it a compelling target for those who would do harm to companies, organizations, nations, and more importantly people. With continued adoption of IP networks, IoT applications have already become a target for attacks that will continue to grow in both magnitude and sophistication. The interconnected nature of IoT devices means that every poorly secured device that is connected online potentially affects the security and resilience of the Internet globally. The weakest link defines the overall level of security of the whole infrastructure. This challenge is amplified by other considerations like the mass-scale deployment of homogenous IoT devices, the ability of some devices to automatically connect to other devices, and the likelihood of fielding these devices in unsecure environments.
IoT presents new challenges to network and security architects. Smarter security systems that include managed threat detection, anomaly detection, and predictive analysis need to evolve [13]. There are various challenges to design security solutions in the IoT because of network characteristics e.g., device heterogeneity, resource constraints, unreliable communication links and the distributed nature. In traditional TCP/IP networks, security is built to protect the confidentiality, integrity and availability of network data. It makes the system reliable and protects the system from malicious attacks which can lead to malfunctioning systems and information disclosure. The IoT requires multi-facet security solutions where the communication is secured with confidentiality, integrity, and authentication services; the network is protected against intrusions and disruptions; and the sensor node as in WSN, additional security protection requirements and user privacy are imposed depending on the application scenario.

With IPv6 there are enough IP addresses to connect billions of ‘things’ to form our new IoT world but whether these things would be secured enough to ensure individual privacy rights and secure the systems from malicious attacks? The cryptographic algorithms are required to be highly efficient, low power, low energy realizations especially for battery operated or passively powered devices. In many practical applications, the gateway needs to send periodic control messages, notifications, and sensitive confidential data to all the wearable devices where a common secret key is required to encrypt the broadcast messages. Symmetric key cryptography such as AES provides fast and lightweight encryption/decryption on such smart devices and their integrated hardware supports it as well. However, when this number of connected devices becomes high, exchanging symmetric keys becomes infeasible and the need to have an efficient scalable key establishment protocol becomes critical. Another approach is to distribute keys by asymmetric key cryptography but it requires high computational costs; the main concern for resource-constrained devices [16]. Therefore, conventional security primitives cannot be applied due to the heterogeneous nature of sensors (either implanted, on-body or wearable), low resources and the system architecture of IoT based healthcare systems [3].

4. IOT SECURITY PRIMITIVES

Devices will only be smart if they include technology to provide security and privacy. Poorly secured IoT devices could serve as entry points for cyber-attack by allowing malicious

individuals to re-program a device or cause it to malfunction. Moreover, unique to cryptographic implementations is that they also need protection against physical tampering either active or passive. This means that countermeasures need to be included during the design process. Security in the IoT must ensure secrecy and integrity of communication, as well as the authenticity of messages being exchanged.

From the end-user's perspective, it is not possible to easily modify these smart devices; security primitives must be pre-embedded into the system. The integration of sensors in the Internet must ensure the interoperability, transparency and flexibility. However, sensor nodes inherently have constrained resources; small batteries are typically the main energy sources for these sensor nodes with the requirement to operate for longer durations [17]. Hence, energy efficiency becomes an important factor besides security and privacy issues.

Different approaches are being employed for secure E2E communication in WSNs and IoT, they can be classified into major research directions as follows

- Centralized Approaches
- Protocol-based Extensions and Optimizations
- Alternative Delegation Architectures
- Solutions that Require Special Purpose Hardware Modules

It is also important to understand the attack techniques in order to rationalize security mechanisms in communication protocols. Some important attacks with respect to IoT are

- Eavesdropping: process of overhearing an ongoing communication, that is as well preliminary for launching next attacks. In wireless communication, everyone has in general access to the medium so takes less effort to launch as compared to wired communication. Confidentiality is a typical counter-measurement against eavesdropping but if keying material is not exchanged in secure manner, eavesdropper could compromise the confidentiality. Secure key exchange algorithms such as Diffe-Hellman (DH) are used.

- Impersonation: a malicious party pretends to be a legitimate entity for instance by replaying a generic message, in order to bypass the aforementioned security goals.

- MITM Attack: Man-in-the-middle attack takes place when a malicious entity is on the network path of two genuine entities. Capable of delaying, modifying or dropping messages. Interesting within the context of PKC, malicious entity doesn't attempt to break the keys of involved parties but rather to become the falsely trusted MITM.

- DoS Attack: targets the availability of a system that offers services, is achieved by exhaustingly consuming resources at the victim so that the offered services become unavailable to legitimate entities. A common way to launch this attack is to trigger expensive operations at the victim that consume resources such as computational power, memory bandwidth or energy. This attack is critical for constrained devices where existing resources are already scarce.

Conventional security primitives cannot be applied due to the heterogeneous nature of sensors, low resources and the system architecture of IoT based systems. Any unauthorized use of data or privacy concerns may restrict people to utilize IoT-based applications. To mitigate these security and privacy threats, strong network security infrastructures are required. Peer authentication and

End-to-End data protection are crucial requirements to prevent eavesdropping on sensitive data or malicious triggering of harmful actuating tasks [15].

Some other security primitives worth consideration are:

- Securing device identity and mechanisms to authenticate it. IoT devices may not have the required compute power, memory or storage to support the current authentication protocols. Therefore, authentication and authorization will require appropriate re-engineering to accommodate our new IoT connected world.

- Protection of the initial configuration and provisioning of devices from tampering, theft and other forms of compromise throughout its usable life, which in many cases can be years.

- Application of geographic location and privacy levels to data

- Strong identities

- Strengthening of other network-centric methods such as the Domain Name System (DNS) with DNSSEC and the DHCP to prevent attacks

- Adoption of other protocols that are more tolerant to delay or transient connectivity (such as Delay Tolerant Networks)

- Lastly, the communication and the data transport channels should be secured to allow devices to send and collect data to and from the agents and the data collection systems. While not all IoT endpoints may have bi-directional communications, leveraging SMS (automatically or via a network administrator) allows secure communication with the device when an action needs to be taken [12].

Data Privacy is another important challenge; privacy fears stemming from the potential misuse of IoT data have captured public attention. There are various questions to be answered: Who owns the data? How a user can be sure that this data is safe and will not be used without his consent? How personal data can be disclosed and used by authorized parties?

Privacy issues are particularly relevant in healthcare, and there are many interesting healthcare applications that fall within the realm of IoT. We can cite among others the tracking of medical equipment in a hospital, the monitoring of vital statistics for patients at home or in an assisted living facility. The system application might be looking for a continuous monitoring of the person's health parameters, while the sensor's ability to record data might limit the sophistication of the security solution used to protect the data it records. In situations for instance, emergency data should be readily available to the medical care unit or responders even without the user's interaction. There would always be a trade-off between functionality and privacy. The one important question still remains: how much privacy are we happy to give up for the potential benefits this new technology can do for humans?

5. IS INTERNET OF THINGS REAL?

Internet of Things is coming. It's not a matter of if or whether, but when and how. And also where do the humans will be placed into this exponentially expanding growth of IoT? Innovations in technology mostly emerge from the needs of human society. Today's top emerging technology:

IoT, focused on proficiently monitoring and controlling different activities will have the impact on human society including everyday life of common people. Ultimately, people will become a part of the IoT through devices for instance in the case of medical implants, without even knowing that they have become part of today's technology.

Here, I will like to quote Mark Weiser's statement in his well-known Scientific American paper, back in 1991. *"The most profound technologies are those that disappear. They weave themselves into the fabric of everyday life until they are indistinguishable from it"*.

6. CONCLUSIONS

This paper aims to give an introductory overview to the reader how wireless sensor are integrated into the Internet of Things, what are the security challenges and the security primitives that might be taken to protect the sensors' data. Current approaches are focused on pre-deployed pre-shared keys on both ends whereas certificate-based authentication is generally considered infeasible for constrained resource sensors. Any unauthorized use of data or privacy concerns may restrict people to utilize IoT-based applications. Peer authentication and data protection are crucial requirements to prevent eavesdropping on sensitive data or malicious triggering of harmful actuating tasks. There are other challenges to be solved if the sensor nodes are integrated into the internet infrastructure and the complete integration of sensor networks and the internet still remains as an open issue. Secret key distribution for heterogeneous sensors in Internet of Things becomes challenging due to the inconsistencies in their cryptographic primitives and computational resources in varying applications. Highly constrained sensors cannot provide enough resources required for the heavy computational operations. The paper studies the interactions between sensor networks and the internet from the point of view of security identifying both the security challenges and the primitives as well.

REFERENCES

[1] Nacer Khalil, Mohamed Riduan Abid, Driss Benhaddou, Michael Gerndt, (2014) "Wireless Sensors Networks for Internet of Things", IEEE Ninth International Conference on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP) Symposium on Public IoT

[2] Jameson Berkow (2011) "Is the Internet leaving humanity behind?" Financial Post

[3] Muhammad A. Iqbal, Dr. Magdy Bayoumi (2016) "Secure End-to-End Key Establishment Protocol for Resource-Constrained Healthcare Sensors in the Context of IoT" The 2016 IEEE International Conference on High Performance Computing and Simulation (HPCS 2016) Innsbruck, Austria

[4] "Internet of Things: An overview by Internet Society" https://www.internetsociety.org/sites/default/files/ISOC-IoT-Overview-20151014_0.pdf

[5] Kashyap Kompella, "A Guide to the Internet of Things"

[6] Azmi Jafarey, "The Internet of Things & IP Address Needs" Network Computing

[7] Phillip Howard, (January 2015) "The Internet of Things Reference Model" Bloor

[8] Therese Sullivan, (November 2014) "The Cutting-Edge of IoT, how does the IoT really change the future of commercial building operations?" Automated buildings

[9] Jim Duffy, (January 2016) "AT&T allies with Cisco, IBM, Intel for city IoT" Network World

[10] Bruce Ndibanje, Hoon-Jae Lee, and Sang-Gon Lee, "Security Analysis and Improvements of Authentication and Access Control in the Internet of Things"

[11] Cristina Alcaraz, Pablo Najera, Javier Lopez, Rodrigo Roman, "Wireless Sensor Networks and the Internet of Things: Do We Need a Complete Integration?" University of Malaga, Spain

[12] Securing the Internet of Things: A Proposed Framework by Cisco

[13] White Paper Internet of Things Intel Corporation (2014). "Developing Solutions for the Internet of Things"

[14] Intel corporations, USA. Intel® Gateway Solutions for the Internet of Things

[15] D. E Vans, (2011)"The Internet of Things: How the Next Evolution of the Internet is Changing Everything", Cisco Internet Business Solutions Group (IBSG).

[16] H. Shafagh and A. Hithnawi, "Poster Abstract: Security Comes First, A Public-key Cryptography Framework for the Internet of Things", 2014 IEEE International Conference on Distributed Computing in Sensor Systems, (2014), pp. 135-136.

[17] Muhammad A. Iqbal, Dr. Magdy Bayoumi, (2016) "A Novel Authentication and Key Agreement Protocol for Internet of Things Based Resource-constrained Body Area Sensors" The IEEE 4th International Conference on Future Internet of Things and Cloud (FiCloud) 2016 Vienna, Austria

AUTHORS

Muhammad A. Iqbal is a graduate student in Computer Science at the University of Louisiana at Lafayette, LA 70504 USA. His research and thesis center around security and privacy for Internet of Things, Wireless Sensor Networks especially in the area of Body Area Networks and healthcare applications in the context of Internet of Things (IoT).

Dr. Magdy A. Bayoumi is the Z.L. Loflin Eminent Scholar Endowed Chair Professor in Computer Science. Dr. Bayoumi has been a faculty member in CACS since 1985. He is the recipient of the 2009 IEEE Circuits and Systems Meritorious Service Award. Dr. Bayoumi is the recipient of the IEEE Circuits and Systems Society 2003 Education Award, and he is an IEEE Fellow. He was on the governor's commission for developing a comprehensive energy policy for the State of Louisiana. He represented the CAS Society on the IEEE National Committee on Engineering R&D policy, IEEE National Committee on Communication and Information Policy, and IEEE National Committee on Energy Policy. He is also active in the "Renewable & Green Energy" and "Globalization: Technology, Economic and Culture" fields. He was a freelance columnist for Lafayette's newspaper.

Dr. Bayoumi has graduated more than 45 Ph.D. and about 175 Master's students. He has published over 300 papers in related journals and conferences. He edited, co-edited and co-authored 5 books in his research interests. He was and has been Guest Editor (or Co-Guest Editor) of eight special issues in VLSI Signal Processing, Learning on Silicon, Multimedia Architecture, Digital and Computational Video, Perception on a Chip, and Systems on a Chip. He has given numerous invited lectures and talks nationally and internationally, and has consulted in industry.

Dr. Bayoumi is the Vice President for Conferences of the IEEE Circuits and Systems (CAS) Society, served in many editorial, administrative, and leadership capacities, including Vice president for technical Activities. He is a technology columnist and writer of the Lafayette newspapers as well.

AUTHOR INDEX