# B V RAJU COLLEGE

VISHNUPUR

BHIMAVARAM

# STUDENT RESEARCH PUBLICATIONS

# ACADAMIC YEAR 2021-22

# INDEX

# NORMALIZATION OF DUPLICATE RECORDS FROM MULTIPLE SOURCES

**A.V.D. SURYA RAO**
PG Scholar, Department of Computer Science,
B.V. RAJU College,
Bhimavaram, W.G.Dt., A.P, India

**Dr. I R KRISHNAM RAJU**
Principal & Professor in Computer Science,
B.V. RAJU College,
Bhimavaram,W.G.Dt., A.P, India

**V.BHASKAR MURTHY**
HOD & Associate Professor in Computer Science,
B.V. RAJU College,
Bhimavaram, W.G.Dt., A.P, India

**Abstract:**

Data consolidation is a challenging issue in data integration. The usefulness of data increases when it is linked and fused with other data from numerous (Web) sources. The promise of Big Data hinges upon addressing several big data integration challenges, such as record linkage at scale, real-time data fusion, and integrating Deep Web. Although much work has been conducted on these problems, there is limited work on creating a uniform, standard record from a group of records corresponding to the same real-world entity. We refer to this task as record normalization. Such a record representation, coined normalized record, is important for both front-end and back-end applications. In this paper, we formalize the record normalization problem, present in-depth analysis of normalization granularity levels (e.g., record, field, and value-component) and of normalization forms (e.g., typical versus complete). We propose a comprehensive framework for computing the normalized record. The proposed framework includes a suit of record normalization methods, from naive ones, which use only the information gathered from records themselves, to complex strategies, which globally mine a group of duplicate records before selecting a value for an attribute of a normalized record. We conducted extensive empirical studies with all the proposed methods. We indicate the weaknesses and strengths of each of them and recommend the ones to be used in practice.

**Keywords :** Data integration, Standards, Task analysis, Databases, Google, Data mining, Terminology

## 1. INTRODUCTION

### 1.1 Introduction:

The usefulness of Web data increases exponentially (e.g., building knowledge bases, Web-scale data analytics) when it is linked across numerous sources. Structured data on the Web resides in Web databases and Web tables. Web data integration is an important component of many applications collecting data from Web databases, such as Web data warehousing (e.g., Google and Bing Shopping; Google Scholar), data aggregation (e.g., product and service reviews), and met searching.

Integration systems at Web scale need to automatically match records from different sources that refer to the same real-world entity find the true matching records among them and turn this set of records into a standard record for the consumption of users or other applications. There is a large body of work on the record matching problem and the truth discovery problem. The record matching problem is also referred to as duplicate record detection, record linkage, object identification, entity resolution, or de-duplication and the truth discovery problem is also called as truth finding or fact finding - a key problem in data fusion.

This work assumes that the tasks of record matching and truth discovery have been performed and that the groups of true matching records have thus been identified. Our goal is to generate a uniform, standard record for each group of true matching records for end-user consumption. It calls the generated record the normalized record. It call the problem of computing the normalized record for a group of matching records the record normalization problem (RNP), and it is the focus of this work.

RNP is another specific interesting problem in data fusion. Record normalization is important in many application domains. For example, in the research publication domain, although the integrator website, such as Citeseer or Google Scholar, contains records gathered from a variety of sources using automated extraction techniques, it must display a normalized record to users. Otherwise, it is unclear what can be presented to users: (i) present the entire group of matching records or (ii) simply present some random record from the group, to just name a couple of ad-hoc approaches. Either of these choices can lead to a frustrating experience for a user, because in (i) the user needs to sort/browse through a potentially large number of duplicate records, and in (ii) it run the risk of presenting a record with missing or incorrect pieces of data. Record normalization is a challenging problem because different Web sources may represent the attribute values of an entity in different ways or even provide conflicting data. Conflicting data may occur because of incomplete data, different data representations, missing attribute values, and even erroneous data.

This work aims to develop a framework for constructing normalized records systematically. This work includes a suit of record normalization methods, from naive ones, which use only the information gathered from records themselves, to complex strategies, which globally mine a group of duplicate records before selecting a value for an attribute of a normalized record.

**1.2 Purpose:**

Record normalization is a challenging problem because different Web sources may represent the attribute values of an entity in different ways or even provide conflicting data. Conflicting data may occur because of incomplete data, different data representations, missing attribute values, and even erroneous data. For example, Table 1 contains four records corresponding to the same entity (publication). They are extracted from different websites. Record Rnorm is constructed by hand for illustration purposes. One notices that the same publication has different representations in different websites.

**1.3 Scope:**

Integration systems at Web scale need to automatically match records from different sources that refer to the same real-world entity find the true matching records among them and turn this set of records into a standard record for the consumption of users or other applications. There is a large body of work on the record matching problem and the truth discovery problem. The record matching problem is also referred to as duplicate record detection, record linkage , object identification, entity resolution ,or deduplication and the truth discovery problem is also called as discovery have been performed and that the groups of true matching records have thus been identified. Our goal is to generate a uniform, standard record for each group of true matching records for end-user consumption. We call the generated record the normalized record. We call the problem of computing the normalized record for a group of matching records the record normalization problem (RNP), and it is the focus of this work. RNP is another specific interesting problem in data fusion.

**1.4 Motivation:**

Record normalization is important in many application domains. For example, in the research publication domain, although the integrator website, such as Citeseer or Google Scholar, contains records gathered from a variety of sources using automated extraction techniques, it must display a normalized record to users. Otherwise, it is unclear what can be presented to users: (i) present the entire group of matching records or(ii)simply present some random record from the group, to just name a couple of ad-hoc approaches. Either of these choices can lead to a frustrating experience for a user, because in (i) the user needs to sort/browse through a potentially large number of duplicate records, and in (ii) we run the risk of presenting a record with missing or incorrect pieces of data.

**1.5 Overview:**

We identify three levels of normalization granularity: record, field, and value-component. Record level assumes that the values of the fields within a record are governed by some hidden criterion and that together create a cohesive unit that is user-friendly. As a consequence, this normalization favors building the normalized record from entire records among the set of matching records rather than piecing it together from field values of different records.Thus,any of the matching records(ideally, that has no missing values) can be the normalized record. Using our running example in Table 1, the record Rc is a possible choice for the normalized record with this level of normalization granularity. Field level assumes that record level is often inadequate in practice because records contain fields with incomplete values. Recall that these records are the products of automatic data extraction tools, which are not perfect and thus may produce errors [18]. This normalization level ignores the cohesion factor in the record normalization level and assumes that a user is better served when each field of the normalized record has as easy to understand a value as possible, selected from among the values in the set of matching records.

## 2. RELATED WORK

SanghyeonBaeg [1] 2008, Power consumption is the most critical issue for low-power ternary content-addressable memory (TCAM) in match lines designs. In the proposed match-line architecture, the match line present in each TCAM word is partitioned into four segments and is selectively pre-charged to reduce the match-line power consumption. The match lines which are partially charged are evaluated to determine the final comparison result by sharing the charges deposited in various parts of the partitioned segments.

B. Heller et al, [2] 2010, Built ElasticTree, which through data-center-wide traffic management and control, introduces energy proportionality in today's non-energy proportional networks. They will likely essentially decrease this quickly developing vitality cost. Compare multiple strategies for finding the minimum-power network [20]. The framework is vitality proficiency, best execution, and adaptation to non-critical failure. The system worked near its ability will build the possibility of dropped and postponed bundles.

A.R. Curtis et al, [3] 2011, DevoFlow proposition enables administrators to target just the streams that issue for their administration issue. DevoFlow handles most miniaturized scale streams in the information plane and consequently enables us to make the most out of switch resources. DevoFlow takes care of the issue by permitting a clonabletrump card principle to choose a yield port. Multipath steering to statically stack balance movement with no utilization of the control-plane. These procedures don't spare much vitality on elite systems.

P. Porraset al, [4] 2012, Incorporates several critical components that are necessary for enabling security applications in Open Flow networks including role-based authorization, rule reduction, conflict evaluation, and policy synchronization. FortNOX is a critical initial phase in enhancing the security of Open Flow systems. It shows the achievability and suitability of our nom de plume set guideline decrease approach [18]. It is unable to handle the dynamic matching process.

Zahid Ullah et al, [5] 2012, Hybrid partitioned static random is a memory architecture in which access memory-based ternary content addressable memory (HP SRAM-based TCAM), which involves TCAM functionality with conventional SRAM, where we are eliminating the inherited disadvantages of conventional TCAMs. HP SRAM-based TCAM is a technique in which they logically dissect conventional TCAM table in a hybrid way (column-wise and row-wise) into TCAM sub-tables, which are then processed to be mapped to their corresponding SRAM memory units.

H. Kim and N. Feamster et al, [6] 2013, Designed and implemented Procera, an event-driven network control framework based on SDN. Additionally, utilize the OpenFlow convention to impart between the Procera controller and the hidden system switches. It gives better permeability and command over undertakings for performing system. This SDN can improve common network management tasks [19]. Procera experiences the characteristic deferral caused by the communication of the control plane and the information plane.

M. Yu, L. Jose et al, [7] 2013, OpenSketch empowers a straightforward and proficient approach to gather estimation information. It utilizes information plane estimation natives dependent on ware switches and an adaptable control plane so administrators can without much of a stretch execute variable estimation calculations. It has a simple, efficient way to control switches [16]. Sketches more flexible in supporting various measurement tasks. Delay of each measurement pipeline component is large.

Weirong Jiang et al, [8] 2013, Random access memory i.e. (RAM)-based Ternary Content Addressable Memory i.e.(TCAM) architecture is design for efficient implementation on state-of-the-art FPGAs. We give a formal study on RAM-based TCAM to disclose the ideas and the algorithms behind it. To face the timing challenge, we propose a modular architecture consisting of arrays of small-size RAM-based TCAM units.

Jacobson et al, [9] 2014, Novel control plane architecture called OpenNF that addresses these challenges through careful API design. OpenNF enables applications to settle on reasonable decisions in meeting their destinations. NF software is always Up-to-Date. The system has High performance on network monitoring.

M. Moshref et al, [10] 2014, DREAM enables operators and cloud tenants to flexibly specify their measurement tasks in a network and dynamically allocates TCAM resources to these tasks based on the resource-accuracy. User-specified high level of accuracy. DREAM can support more concurrent tasks. DREAM needs to dismiss almost half of the assignments and drop about 10%.

N. Katta et al, [11] 2014, CacheFlow system is a system which “caches” the most popular rules in the small TCAM, in which they are relying on software to handle the small amount of “cache miss” traffic. But, we cannot blindly apply existing cache-replacement algorithms, because of dependencies between rules with overlapping patterns.

Naga Katta et al, [12] 2014, Instead of creating long dependency chains to cache smaller groups of rules in which semantics of the network policy are preserved. There are mainly four types of criteria for it. Elasticity which combines the best of hardware and software switches. Transparency which faith-fully supporting native OpenFlow semantics, including traffic counters. Fine-grained rule caching which places popular rules in the TCAM, despite dependencies on less-popular rules. Adaptability which enables incremental changes to the rule caching as the policy changes.

## 3. EXISTING SYSTEM

In existing, TCAM Razor, DomainFlow and Palette algorithms are used. Which is both power hungry and highly limited in capacity . Most TCAM-capable commodity switches support only a few thousand wildcard entries. Although certain products recently reported an ability to support up to 125k wildcard entries, enlarging the memory with enhanced control capability significantly increases the cost To improve scalability, two approaches have been taken: proactively allocating rules on multiple switches to load balance the memory consumption , and reactively caching rules on each switch individually.

**3.1 Disadvantages**: These existing works provides poor caching ratio and less hit ratioThese existing works provides poor caching ratio and less hit ratio

## 4. PROPOSED SYSTEM

To deal with existing disadvantages, this work proposed a novel wildcard-rule caching algorithm and a cache replacement algorithm to make use of TCAM space efficiently.TCAM can look up a packet's header and compare the matching patterns of the packet to the match field of all rules in the flow table in parallel. Our wildcard-rule caching algorithm repeats caching a set of important rules into TCAM until there is no TCAM space. Our cache replacement algorithm takes temporal and spatial traffic localities into consideration, which could make hit ratio high.

### 4.1 Advantages

The proposed wildcard-rule caching algorithm could have better caching ability than the other existing algorithms. Furthermore, the proposed cache replacement algorithm could have higher hit ratio than the other existing algorithms.

## 5. IMPLEMENTATION

### 5.1 Load conference name dataset:

This module load conference name dataset.This dataset contain rid, label and conference name.This dataset contains 3683 records.

### 5.2 Mining Abbreviation Definition pairs:

This module use a number of heuristics to determine whether given two value components s and t, s is an abbreviation of t. In this section, a value component is a word (or term). As we mentioned previously, in this module we consider only fields with the string data type. We define the neighboring context of a word w within the set of values of a field fj as the set of pairs (left neighbor word, right neighbor word) with the property that the substring left neighbor word w right neighbor word is a substring of a value of fj in some record in Re. If w is the beginning word of a field value, we use a special start-symbol "<s>" to mark left neighbor word. If it is the last word in the field value, we use the special end-symbol "</s>" to mark right neighbor word. For example, the words "proceedings" and "proc" occur many times in the field venue, and they share a good fraction of their neighboring contexts, such as (in, of), (<s>, of), (in, acm). "proc" is also the prefix of "proceedings", so we become increasingly

confident that "proc" is a possible abbreviation of "proceedings".

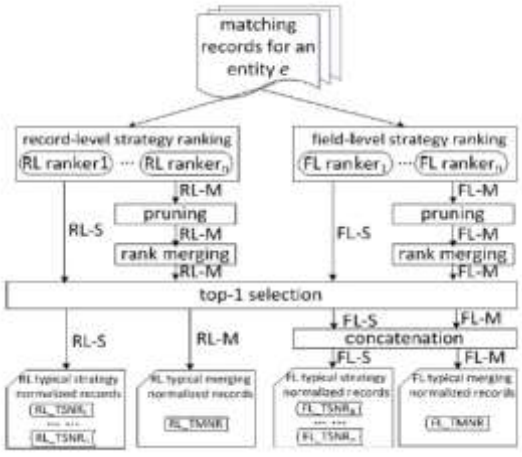
### 5.3 Mining TemplateCollocation-SubCollocation Pairs (MTS):

This module aim to find all template collocations and their subcollocations. The template collocations become the candidates with which it can expand (replace) the subcollocations. They will be used to generate the normalized component values for a field. Let an n-collocation tc be a template collocation and a k collocation kc be its subcollocation (k < n).

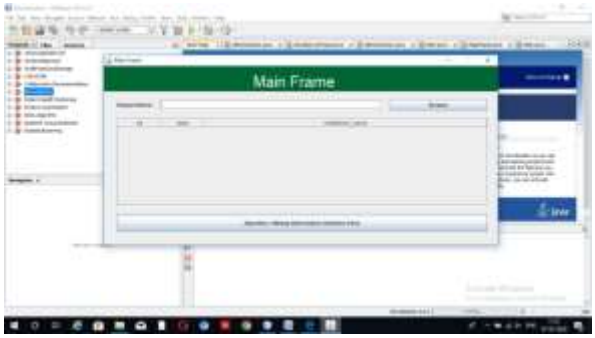
### 5.4 Mining Most Frequently Co-occurring Template Collocation:

The above module, discussed how to obtain the template collocations and their corresponding subcollocations. We notice that some of the template collocations co-occur frequently. For example, among the values of the field venue, the template collocation "conference on" co-occurs most frequently with "in proceedings of the." We also observe that template collocation co-occurrence is not always bidirectional. For example, the template collocation "symposium on" co-occurs most often with "in proceedings of the", but "in proceedings of the" co-occurs most frequently with "conference on."

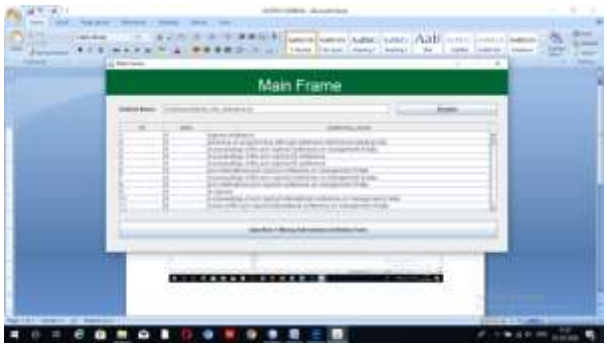
## 6.Architecture



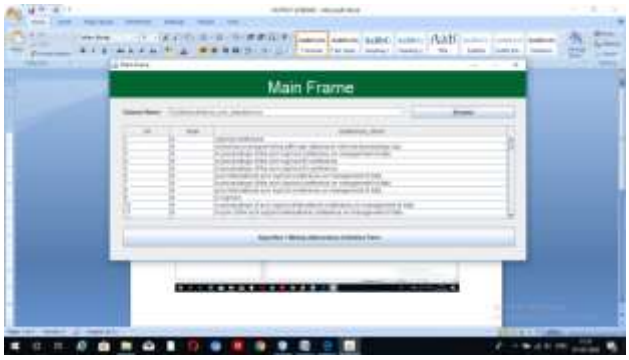

## 7. OUTPUT RESULTS



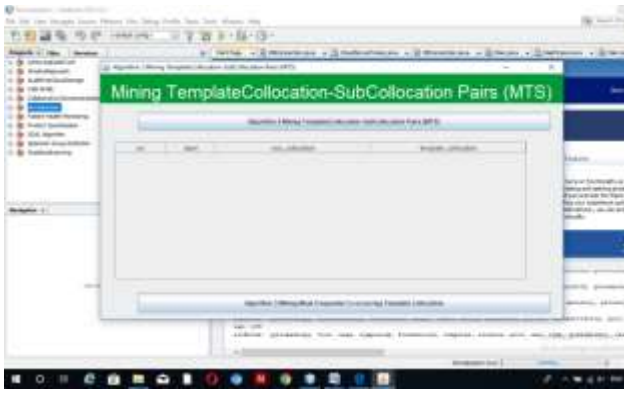

Load Dataset Screen



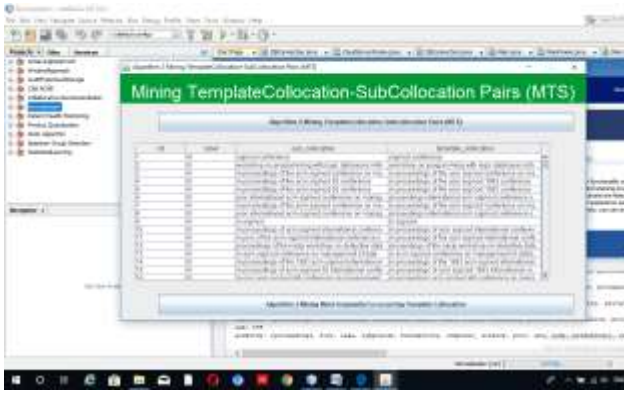

Show Dataset screen



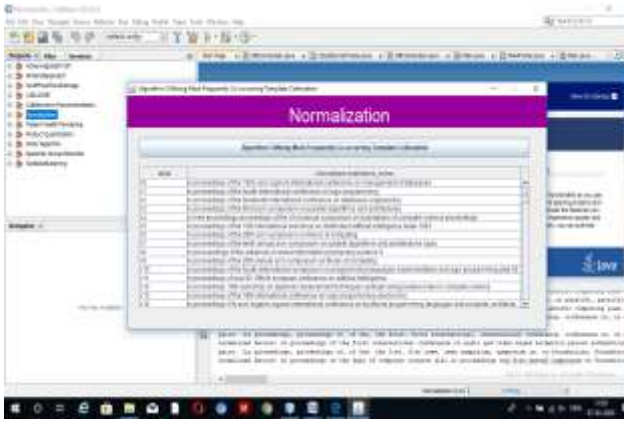

**Implement Algorithm-1**




**Implement Algorithm-2**




Minig Template collection-sub collection pairs screen




**Implement Algorithm-3**

## 8. CONCLUSION AND FUTURE ENHANCEMENT

This work studied the problem of record normalization over a set of matching records that refer to the same real-world entity. This work presented three levels of normalization granularities (record-level, field-level and value component level) and two forms of normalization (typical normalization and complete normalization). For each form of normalization, this work proposed a computational framework that includes both single-strategy and multi-strategy approaches. This work proposed four single-strategy approaches: frequency, length, centroid, and feature-based to select the normalized record or the normalized field value. For multistrategy approach, this work used result merging models inspired from metasearching to combine the results from a number of single strategies. This work analyzed the record and field level normalization in the typical normalization. In the complete normalization, this work focused on field values and proposed algorithms for acronym expansion and value component mining to produce much improved normalized field values. This work implemented a prototype and tested it on a real-world dataset. The experimental results demonstrate the feasibility and effectiveness of this approach. This method outperforms the state-of-the-art by a significant margin**.**

## 9. BIBLIOGRAPHY

[1] A. Culotta, M. Wick, R. Hall, M. Marzilli, and A. McCallum, "Canonicalization of database records using adaptive similarity measures," in SIGKDD, 2007, pp. 201–209.

[2] O. Benjelloun, H. Garcia-Molina, D. Menestrina, Q. Su, S. E. Whang, and J. Widom, "Swoosh: A generic approach to entity resolution," VLDBJ, vol. 18, no. 1, pp. 255–276, 2009.

[3] M. L. Wick, K. Rohanimanesh, K. Schultz, and A. McCallum, "A unified approach for schema matching, coreference and canonicalization," in SIGKDD, 2008, pp. 722–730.

[4] S. Tejada, C. A. Knoblock, and S. Minton, "Learning object identification rules for information integration," Inf. Sys., vol. 26, no. 8, pp. 607–633, 2001.

[5] L. Wang, R. Zhang, C. Sha, X. He, and A. Zhou, "A hybrid framework for product normalization in online shopping," in DASFAA, vol. 7826, 2013, pp. 370–384.

[6] S. Chaturvedi and et al., "Automating pattern discovery for rule based data standardization systems," in ICDE, 2013, pp. 1231– 1241.

[7] E. C. Dragut, C. Yu, and W. Meng, "Meaningful labeling of integrated query interfaces," in VLDB, 2006, pp. 679–690.

[8] S. Raunich and E. Rahm, "Atom: Automatic target-driven ontology merging," in ICDE, 2011, pp. 1276–1279.

# An Efficient Feedback Control Mechanism for Positive or Negative Information Spread in Online Social Networks

**A. N.V.SATYA SAI MANOJ**
PG Scholar, Department of Computer Science,
B.V. RAJU College,
Bhimavaram, W.G.Dt., A.P, India

**Dr. I R KRISHNAM RAJU**
Principal & Professor in Computer Science,
B.V. RAJU College,
Bhimavaram,W.G.Dt., A.P, India

**V. BHASKAR MURTHY**
HOD & Associate Professor in Computer Science,
B.V. RAJU College,
Bhimavaram, W.G.Dt., A.P, India

***Abstract-*** The wide availability of online social networks (OSNs) facilitates positive information spread and sharing. However, the high autonomy and openness of the OSNs also allow for the rapid spread of negative information, such as unsubstantiated rumors and other forms of misinformation that often elicit widespread public cognitive misleads and huge economic losses. Therefore, how to effectively control the negative information spread accompanied by positive information has emerged as a challenging issue. Unfortunately, this issue still remains largely unexplored to date. To fill this gap, we propose an efficient feedback control mechanism for the simultaneous spread of the positive and negative information in OSNs. Specifically, a novel computational model is first proposed to present the temporal dynamics of the positive and negative information spread. Furthermore, the proposed mechanism restrains the negative information spread with minimal system expenses by devising and performing three synergetic intervention strategies. Technically, this mechanism intensively evaluates the number of seed users performing three intervention strategies. Besides, each seed user performs the received control task independently, and then the control plan for the next time step is adjusted dynamically according to the previous feedback results. Finally, we evaluate the efficiency of the proposed mechanism based on the extensive experimental results obtained from two real-world networks.

***Index Terms:*** Integrated circuit modeling, Feedback control, Biological system modeling, Social networking (online), Epidemics, Computational modeling, Mathematical model

## I. Introduction

THE INCREASING popularity of online social networks (OSNs), such as Facebook,1 Twitter,2 and LinkedIn,3 has created a fruitful environment for the spread of positive information [1]–[3]. However, the high openness and autonomy of the OSNs also enable the spread of negative information, such as unsubstantiated rumors, conspiracy theories, and other forms of misinformation [4]–[6]. Specifically, the authentic information is called positive information and the false news (e.g., malicious rumors) is called negative information [7]. Both positive and negative information can spread in the OSNs if users are willing to believe and forward the received positive or negative information. For instance, the misinformation about the *Ebola outbreak* has caused confusion among public healthcare workers who tried to combat the outbreak [8].

Seriously, sowing seeds of disbelief and fear only makes it more challenging to bring the *Ebola outbreak* under control. Among this case, users in the OSNs might have been misled by the

received rumor, that is, they may believe and spread the received negative information and are in the negative information-spreading state. Conversely, once the authority (the positive information source) clarifies the rumor to the public, they turn to believing and spreading the official information and are in the positive information spreading state. Furthermore, when the users receive both the positive and negative information and cannot decide whether to believe the positive or the negative information, they are in the dual information-hesiting state. Therefore, the simultaneous spread of the positive and negative information in OSNs exhibits a complicated spread process. Undoubtedly, how to control the negative information spread accompanied by positive information has become critical issue. A reasonable model for the simultaneous spread of the positive and negative information in OSNs should take the key factors influencing the user's decision making into consideration. For example, when positive and negative information spread in OSNs, some users may believe either the positive information or the negative information, or neither kind of information.

## II. Related Work

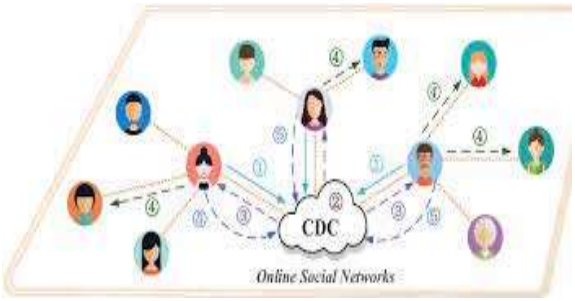
Opinion mining involves several important tasks, including sentiment polarity and intensity assignment [18], [31]. Polarity assignment is concerned with determining whether a text has a positive, negative, or neutral semantic orientation. Sentiment intensity assignment looks at whether the positive/negative sentiments are mild or strong. Given the two phrases "I don't like you" and "I hate you," both would be assigned a negative semantic orientation but the latter would be considered more intense. Effectively classifying sentiment polarities and intensities entails the use of classification methods applied to linguistic features. . popular class of features used for opinion mining is n-grams [28], [38]. Various n-gram categories have attained state-of-the-art results [3], [27]. Larger n-gram feature sets require the use of feature selection methods to extract appropriate attribute subsets. Next, we discuss these two areas: n-gram features and feature selection techniques used for Author profiling. [1], [2], [7], [29]. The most Polarity assignment is concerned with determining whether a text has a positive, negative, or neutral semantic orientation. Sentiment intensity assignment looks at whether the positive/negative sentiments are mild or strong. Given the two phrases "I don't like you" and "I hate you," both would be assigned a negative semantic orientation but the latter would be considered more intense. Effectively classifying sentiment polarities and intensities entails the use of classification methods applied to linguistic features.

## N-GRAM FEATURES FOR AUTHOR PROFILING

N-gram features can be classified into two categories: fixed and variable. Fixed n-grams are exact sequences occurring at either the character or token level. Variable n-grams are extraction patterns capable of representing more sophisticated linguistic phenomena. A plethora of fixed and variable n-grams have been used for opinion mining, including word, part-of-speech (POS), character, legomena, syntactic, and semantic n-grams. Word n-grams include bag-of-words (BOWs) and higher order word n-grams (e.g., bigrams, trigrams). Word n-grams have been used effectively in several studies [28]. Typically, unigrams to trigrams are used [3], [27], though 4- grams have also been employed [34]. Word n-grams often provide a feature set foundation, with additional feature categories added to

them [4], [27], [34], [38]. Given the pervasiveness of adjectives and adverbs in opinion-rich text, POS tag, n-grams are very useful for sentiment classification [10], [12]. Additionally, some studies have employed word plus part-of-speech (POS Word) n-grams. These n-grams consider a word along with its POS tag in order to overcome word-sense disambiguation in situations where a word may otherwise have several senses [38]. For example, the phrase "quality of the" can be represented with the POS Word trigram "quality-noun of prep the-det." Character n-grams are letter sequences. For example, the word "like" can be represented with the following two and three letter sequences "li, ik, ke, lik, ike." While character n-grams were previously used mostly for style classification, they have recently been shown to be useful in related affect classification research attempting to identify emotions in text [2]. Legomena n-grams are collocations that replace once (hapax legomena) and twice occurring words (dis legomena) with "HAPAX" and "DIS" tags [2], [38].

**Architecture**




## III. EXISTING SYSTEM

- Researchers have studied the temporal dynamics of a single type of information, such as rumor, virus, etc. [12], [18], [19]. Recently, there has been an increasing interest in developing efficient methods for information spread minimization by using biological epidemic models. Chen *et al.* [18] employed the epidemic model to describe the collective dynamics of information spread over networks. They provided an analytical model for information spread to solve the optimal control signal distribution time for minimizing the total network cost via dynamic programming. He *et al.* [19] developed a heterogeneous network-based epidemic model that incorporates the rumor blocking and truth spreading strategies to characterize the rumor spreading in mobile social networks (MSNs). The rumor blocking strategy was conducted by optimally combining various rumor restraining methods such that a rumor can be extinct within an expected time period. The truth spreading strategy was implemented by spreading truth periodically.
- Ji *et al.* [24] investigated the problem of identifying multiple rumor or infection sources which may start spreading at different times under the susceptible–infected (SI) model [25]. They introduced a quasi-regular tree and a heavy center to design an algorithmic framework which can transform an abstract estimator into a two-source joint estimator. Jiang *et al.* [26] identified rumor sources from time-varying social networks. They simplified the time-varying networks as a series of static networks and employed a reverse spread strategy to detect a set of suspects of the real rumor source. In addition, they introduced a novel rumor spreading model to calculate the maximum likelihood for each suspect to determine the real source from the suspects. On the other hand, in order to minimize the influence of the rumor spread by blocking a certain of the subset of nodes, Wang *et al.* [27] presented a model of dynamic rumor

influence minimization with a user experience which considers both the global popularity and individual attraction of the rumor.

**Disadvantages**

- In the existing work, the system is very less effective due to Spread of negative information.
- The system doesn't have ability to resist rumor spreading in the OSN.

## IV. PROPOSED SYSTEM

The system proposes an efficient feedback control mechanism for simultaneous positive and negative information spread in OSNs. Specifically, we present a novel computational model to describe the temporal dynamics of the positive and negative information spread. Then, three synergetic intervention strategies for users in the unknown state, negative information-spreading state, and dual information hesitating state are developed, respectively. Furthermore, the proposed mechanism restrains the negative information spread by performing three synergetic intervention strategies with minimal system expenses, which is characterized by centralized computing, distributed execution, and dynamic optimization. The main contributions of this article are summarized as follows.

❖ The system analyzes the trust relationships among users and quantifies the transfer probability between user's behavioral states (i.e., unknown state, positive information spreading state, negative information-spreading state, and dual information-hesitating state in the OSNs with positive and negative information spread). Then, a novel computational model is established based on the differential equations to characterize the temporal dynamics under the coexistence of the positive and negative information.

❖ In order to ensure that the negative information can be controlled more efficiently, we develop three synergetic intervention strategies (i.e., warning, correction, and guidance) for users in the unknown state, negative information-spreading state, and dual information hesitating state, respectively. Then, the minimization of the total system expenses is modeled as an optimal control problem. Moreover, we utilize Pontryagin's maximum principle to obtain the dynamic distribution of the three synergetic intervention strategies over time.

❖ The system devises a nonlinear feedback control mechanism (NFCM) to perform three synergetic intervention strategies with minimal system expenses. Specifically, NFCM first evaluates the number of seed users who perform three strategies at each time step intensively. Then, each seed user performs the control task independently. Third, NFCM adjusts the control plan for the next time step dynamically based on the previous feedback results. The key originality of this article is to pioneer an efficient feedback control mechanism for the simultaneous spread of the positive and negative information in OSNs.

❖ The system evaluates the performance of the proposed feedback control mechanism based on two real-world datasets. The experimental results demonstrate that the proposed

❖ NFCM can restrain the negative information spread and decrease the total system expenses by 73% compared to the case without the control mechanism proposed in this article. In addition, further experimental results show that the total system expenses are decreased if the proposed mechanism performs intervention strategies earlier.

**Advantages**

❖ An efficient feedback control mechanism for simultaneous positive

and negative information spread in OSNs.

- ❖ The system devises a nonlinear feedback control mechanism (NFCM) to perform three synergetic intervention strategies with minimal system expenses.

## V.IMPLEMENTATION

- **Admin** In this module, the Admin has to login by using valid user name and password. After login successful he can perform some operations such as View All Users and Authorize, Add Filters, View All Posts, View Friend Request and Response,View All Positive Feedback,View All Negative Feedback,View All Feedback Results.
- **Friend Request & Response**
  In this module, the admin can view all the friend requests and responses. Here all the requests and responses will be displayed with their tags such as Id, requested user photo, requested user name, user name request to, status and time & date. If the user accepts the request then the status will be changed to accepted or else the status will remains as waiting.
  **Social Network Friends**
  In this module, the admin can see all the friends who are all belongs to the same site. The details such as, Request From, Requested user's site, Request To Name, Request To user's site.

## VI.CONCLUSION

In this article, we have addressed the problem of the coupling spread of the positive and negative information in OSNs. First, we have established a coupling spread model of the positive and negative information to describe the dynamic coupling spread process. In addition, we have proposed three synergetic control strategies to control the coupling spread process of positive and negative information. Afterward, we have devised an NFCM to perform three synergetic control strategies with minimal system expenses. The experimental results demonstrate that our proposed NFCM can effectively decrease the spread of negative information.

Future directions include the investigation of the network topology and node property on the spread process of positive and negative information. Then, we intend to employ the idea of crowd sourcing to design a distributed control algorithm to intervene in the coupling spread of the positive and negative information.

## REFERENCES

[1] C.-K. Chou and M.-S. Chen, "Learning multiple factors-aware diffusionmodels in social networks," IEEE Trans. Knowl. Data Eng., vol. 30,no. 7, pp. 1268–1281, Jul. 2018.

[2] S. Gao, H. Pang, P. Gallinari, J. Guo, and N. Kato, "A novel embedding method for information diffusion prediction in social network big data,"IEEE Trans. Ind. Inform., vol. 13, no. 4, pp. 2097–2105, Aug. 2017.

[3] Q. Liang and E. Modiano, "Network utility maximization in adversarial environments," in Proc. IEEE Conf. Comput. Commun. (INFCOM), Apr. 2018, pp. 594–602.

[4] M. D. Vicario et al., "The spreading of misinformation online," Proc.Nat. Acad. Sci. USA, vol. 113, no. 3, pp. 554–559, 2016.

[5] S. Zhao, Y. Gao, G. Ding, and T.-S. Chua, "Real-time multimedia social event detection in microblog," IEEE Trans. Cybern., vol. 48, no. 11,pp. 3218–3231, Nov. 2018.

[6] R. Dong, L. Li, Q. Zhang, and G. Cai, "Information diffusion on social media during natural disasters," IEEE Trans. Comput. Soc. Syst., vol. 5,no. 1, pp. 265–276, Mar. 2018.

[7] S. Wen, M. S. Haghighi, C. Chen, Y. Xiang, W. Zhou, and W. Jia, "Asword with two edges: Propagation studies on both positive and negative information in online social networks," IEEE Trans. Comput., vol. 64 no. 3, pp. 640–653, Mar. 2015.

[8] (Oct. 2014). The Inevitable Rise of EBOLA Conspiracy Theories.[Online]. Available: https://www.washingtonpost.com/news/wonk/wp/2014/10/13/the-inevitable-rise-of-ebola-conspiracy-theories

[9] E. Stai, E. Milaiou, V. Karyotis, and S. Papavassiliou, "Temporal dynamics of information diffusion in Twitter: Modeling and experimentation," IEEE Trans. Comput. Soc. Syst., vol. 5, no. 1, pp. 256–264, Mar. 2018.

[10] A. Calió, R. Interdonato, C. Pulice, and A. Tagarelli, "Topology-driven diversity for targeted influence maximization with application to user engagement in social networks," IEEE Trans. Knowl. Data Eng., vol. 30, no. 12, pp. 2421–2434, Dec. 2018.

[11] X. Wang, L. Zhang, Y. Lin, Y. Zhao, and X. Hu, "Computational models and optimal control strategies for emotion contagion in the human population in emergencies," Knowl. Based Syst., vol. 109, pp. 35–47, Oct. 2016.

[12] C. Zheng, C. Xia, Q. Guo, and M. Dehmer model with structural diversity for social networks," IEEE Trans. Cybern., vol. 47, no. 4, pp. 1078–1089, Apr. 2017.

[13] G. Zhang, D. Lu, and H. Liu, "Strategies to utilize the positive emotionalcontagion optimally in crowd evacuation," IEEE Trans. Affect. Comput.,vol. 14, no. 8, pp. 1–14, Jul. 2018.

[14] X. Liu, D. He, and C. Liu, "Information diffusion nonlinear dynamics modeling and evolution analysis in online social network based on emergency events," IEEE Trans. Comput. Soc. Syst., vol. 6, no. 1, pp. 8–19, Feb. 2019.

[15] Q. Bao, W. K. Cheung, Y. Zhang, and J. Liu, "A component-based diffusion.

[16] Y. Cen et al., "Trust relationship prediction in Alibaba e-commerce platform," IEEE Trans. Knowl. Data Eng., early access, doi: 10.1109/TKDE.2019.2893939. , "Interplay between sir-based disease spreading and awareness diffusion on multiplex networks," J. Parallel Distrib. Comput., vol. 115, no. 5, pp. 20–28, May 2018.

[17] Y.-B. Zhuang, J. J. Chen, and Z.-H. Li, "Modeling the cooperative and competitive contagions in online social networks," Physica A Stat. Mech Appl., vol. 484, pp. 141–151, Oct. 2017.

[18] P. Chen, S. Cheng, and K. Chen, "Optimal control of epidemic information dissemination over networks," IEEE Trans. Cybern., vol. 44,no. 12, pp. 2316–2328, Dec. 2014. [19] Z. He, Z. Cai, J. Yu, X. Wang, Y. Sun, and Y. Li, "Cost-efficient strategiesfor restraining rumor spreading in mobile social networks," IEEE Trans. Veh. Technol., vol. 66, no. 3, pp. 2789–2800, Mar. 2017.

[20] D. Xu, X. Xu, Y. Xie, and C. Yang, "Optimal control of anSVRS epidemic spreading model with virus variation based on complexnetworks," Commun. Nonlinear Sci. Numer. Simulat., vol. 48,pp. 200–210, Jul. 2017.

[21] K. Li, G. Zhu, Z. Ma, and L. Chen, "Dynamic stability of an siqsepidemic network and its optimal control," Commun. Nonlinear Sci.Numer. Simulat., vol. 66, pp. 84–95, Jan. 2019.

[22] L. Li, Q. Xu, T. Gan, C. Tan, and J.-W. Lim, "A probabilistic model ofsocial working memory for information retrieval in social

interactions,"IEEE Trans. Cybern., vol. 48, no. 5, pp. 1540–1552, May 2018.

[23] Z. Yu, F. Yi, Q. Lv, and B. Guo, "Identifying on-site users for socialevents: Mobility, content, and social relationship," IEEE Trans. MobileComput., vol. 17, no. 9, pp. 2055–2068, Sep. 2018.

[24] F. Ji, W. P. Tay, and L. R. Varshney, "An algorithmic framework forestimating rumor sources with different start times," IEEE Trans. SignalProcess., vol. 65, no. 10, pp. 2517–2530, May 2017.

[25] K. Liu, T. Zhang, and L. Chen, "State-dependent pulse vaccination andtherapeutic strategy in an Si epidemic model with nonlinear incidencerate," Comput. Math. Metho**d** Med., vol. 2019, no. 2, pp. 1–10, 2019.

# A SURVEY ON OPTICAL CHARACTER RECOGNITION SYSTEM

**B.V.DURGA SAI RAMYA**
PG Scholar, Department of Computer Science,
B.V. RAJU College,
Bhimavaram, W.G.Dt., A.P, India

**Dr. I R KRISHNAM RAJU**
Principal & Professor in Computer Science,
B.V. RAJU College,
Bhimavaram,W.G.Dt., A.P, India

**V.BHASKAR MURTHY**
HOD & Associate Professor in Computer Science,
B.V. RAJU College,
Bhimavaram, W.G.Dt., A.P, India

**ABSTRACT**

At present scenario, there is growing demand for the software system to recognize characters in a computer system when information is scanned through paper documents. This paper presents detailed review in the field of Optical Character Recognition. Various techniques are determined that have been proposed to realize the center of character recognition in an optical character recognition system. OCR (Optical Character Recognition) translates images of typewritten or handwritten characters into the electronically editable format and it preserves font properties. Different techniques for preprocessing and segmentation have been surveyed and discussed in this paper.

**Keywords: Character Recognition System, Image Segmentation, OCR, Preprocessing, Skew correction, Classifier.**

## 1. INTRODUCTION

OCR (Optical Character Recognition) translates images of typewritten or handwritten characters into machine editable format. OCR reads damaged or low-quality codes and returns the best guess at what the code is. It is widely used as a form of information entry from printed paper data records, whether passport documents, invoices, bank statements, computerized receipts, business cards, mail, printouts of static data, or any suitable documentation. OCR does not deal with quality and sharpness of characters. To overcome the limitations of OCR a new approach comes into picture which is OCV. Projection Profile-based methods used makes segmentation easy to separate the text in document image into lines, words, and characters independent of the Language in the Text. Different methods are used at each intermediate stage of OCR. Text Segmentation is done using Projection Profile method. They proposed an algorithm for correction of the skew angle of the text document [1]. Blur is the important factor that damages OCR accuracy. In this paper prediction method based on a local blur estimation is proposed. The relation between blur effect and character size is investigated which is useful for the classifier. Classifier separates the given document into

three classes: readable, intermediate, non-readable classes [2].

The grading system is used to evaluate the performance of printed text using various quality measures. The recognition results showed high recognition rate as the system was able to perform a recognition rate of 98.69 % along with a precision of 0.9857 and a sensitivity of 1 [3]. This paper presents complete OCR (Optical Character Recognition) system for camera captured image/graphics embedded textual documents for handheld devices [4]. Paper [5] describes the skew detection and correction of scanned document images written in Assamese language using the horizontal and vertical projection profile analysis OCR consists of many phases such as Pre-processing, Segmentation, Feature Extraction, Classifications and Recognition [6].

**1.1 Digitization**

Digitization is the process of converting a paper-based handwritten document into electronic format. Here, each document consists of only one character. The electronic conversion is accomplished by using a method whereby a document is scanned and an electronic representation of the original document as an image file format is produced. The author used various scanners for digitization, and the digital image was going for next step that is a preprocessing phase.

**1.2 Pre-processing** In The pre-processing phase, there is a series of operations performed on the scanned input image. It enhances the image rendering it suitable for segmentation the gray-level character image is normalized into a window sized. After noise reduction, a bitmap image is produced. Then, the bitmap image was transformed into a thinned image.

**1.3 Segmentation** The Segmentation phase is the most important process. Segmentation is done by separation from the individual characters of an image. Segmentation of handwritten characters into different zones (upper, middle and lower zone) and characters is more difficult than that of printed documents that are in standard form. This is mainly because of variability in a paragraph, words of line and characters of a word, skew, slant, size and curved. Sometimes components of two adjacent characters may be touched or overlapped and this situation creates difficulties in the segmentation task. The touching or overlapping problem occurs frequently because of modified characters in upper-zone and lower-zone.

**1.4 Feature Extraction and classification** Feature extraction is the phase which is used to measure the relevant shape contained in the character. In the feature extraction phase, one can extract the features according to levels of text, e.g., character level, word level, line level and paragraph level. The classification phase is the decision making phase of an OCR engine, which uses the features extracted in the previous

stage for making the class memberships in pattern recognition system. The preliminary aim of classification phase of OCR is to develop the constraint for reducing the misclassification relevant to feature extractions.

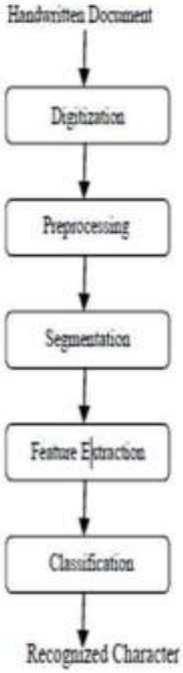



Figure 1 Phases of General Character Recognition System

## 2. METHODOLOGY

One of the most important steps of offline character recognition system is skew detection and correction which has to be used in scanned documents as a pre-processing stage in almost all document analysis and recognition systems. This paper describes the skew detection and correction of scanned document images written in Assamese language using the horizontal and vertical projection profile analysis [5]. Documents with background images in OCR cause an error. A non-linear transformation is used to enhance the contrast of each channel image. The experimental results show that the recognition accuracies are improved significantly after removing background images [7]. For pre-processing Fourier Transform is used which decomposes an image into sine and cosine components with increasing frequencies. Fourier transform converts spatial domain onto frequency domain which is easily used for further processing [1]. Reading text from photographs is a challenging problem. They applied recently developed machine learning algorithms for learning the features automatically from unlabeled data. They proposed text detection and recognition system based on a scalable feature learning algorithm and applied it to images of text in natural scenes [8]. Since past few years, research has been performed to develop machine printed Chinese/English characters. In this paper, they described the search and fast match techniques. High-performance Chinese/English OCR engine is used to construct a large vocabulary. They have collected 1862 text lines from varied sources such as newspapers, magazines, journals, books, etc [9].

H. Wang and J. Kangas [10] proposed a method of identifying character- like regions in order to extract and recognize characters in natural color scene images automatically. Connected component extraction is used to check the block candidates. Priority adaptive segmentation (PAS) is implemented to obtain

accurate foreground pixels of the character in each block. Paper [11] presented a system for text extraction based on the open-source OCR algorithm. The system is used for functional verification of TV sets. J. Diaz-Escobar [12] proposed a new method for recognition of content-less characters in degraded images using the phase congruency and local energy model. The suggested phase features are invariant to non-uniform illumination and slight geometric distortions. Degraded images were compared with that of the SIFT method in terms of recognition metrics. Another approach in the paper [13] Hauling the scene text from image and video is challenging due to the complex background, changeable font size, dissimilar style, unknown layout, poor resolution and blurring, position, viewing angle and so on. For text extraction region and connected component based methods are used. Artificial Neural network (ANN) is used as the classifier to filter out the text and non-text components. There are natural variations in human writing so designing a reliable OCR system is a challenging task. An algorithm based on Kohonen Neural Network is presented in this paper. Kohonen algorithm that is one of Artificial neural network The experiments also demonstrated that system complexity can be reduced significantly without degrading performance by considering two-layered neural network rather than multiple layered neural networks [14]. In this paper [15] a complete OCR methodology for recognizing historical documents, either printed or handwritten without any knowledge of the font, is presented. The pre-processing and segmentation approach is used in order to detect text lines, words, and characters. Yaeger [16] has proposed a handwritten character recognition system. The proposed system works by using the neural network techniques. For the recognition of characters, a multi-layer perceptron is used by this system and it gives better results. J. Hu et. al [17] proposed a system in which high-level features are combined with low-level features on simple points and these are able to cover a huge amount of input patterns. Also, these features have invariance property which is used for normalizing the curvature of features. Funanda [18] has proposed a system which uses the HMM for the recognition of the online handwritten recognition. The proposed system reduces the usage of memory and also it improved the recognition rate of online handwritten characters. In paper [1] Horizontal Projection Profile and Vertical Projection Profile methods are used for segmentation. Different methods are used at each intermediate stage of OCR. Text Segmentation is done using Projection Profile method. They proposed an algorithm for correction of the skew angle of the text document.

J. r´ı Matas [19] presented an end-to-end real-time scene text localization and recognition method. In the first stage of the classification, the probability of each ER being a character is estimated using novel features calculated with

O(1) complexity. In second stage only ERs with locally maximal probability are selected. Huei-Yung Lin and Chin-Yu Hsu [20] presented neural network based approach which reduces the training time and maintains the high recognition rate. Multi-stage approach and pre-processing are done for the experiment. Preprocessing is performed to partition the training data prior to training stage. In this paper [21], a computer vision and character recognition algorithm for a license plate recognition (LPR) is presented to be used as a core for intelligent infrastructure like electronic payment systems (toll payment, parking fee payment), freeway. Based on the connected component analysis and novel adaptive image segmentation technique is presented [21].

## 3. COMPARISON

Paper [5] presented that projection profile is used as a suitable feature for skew detection. Vertical Projection Profile Analysis allows small noise which produces error where Horizontal Projection Profile Analysis reduces the effect of noise. The time complexity of Vertical is high with compared to horizontal projection profile. In paper [7] author proposed a method which is used to remove the background image from pilling up. In government agencies and independent organizations, OCR simplifies data collection and analysis, among other processes, document. The experiment is done using three OCR software tool: HANWANG OCR, ABBYY, and Tesseract. With compared to Tesseract OCR, HANWANG OCR, and ABBYY OCR better because there are built-in functions are available to preprocess image before text extraction In paper [8] they trained their character classifier with features. They tested 5198 characters from 62 classes 26 upper- and 26 lower-case letters). Accuracy for the largest system (1500 features) is the highest, at 81.7% for the 62-way classification problem. Segmentation is an important stage of OCR in image processing. In this paper [22] they surveyed different techniques which are available for segmentation. Most methods are categorized into three groups: the analytical, the empirical goodness and the empirical discrepancy groups. Segmentation algorithms can be evaluated analytically or empirically, so the evaluation methods can be divided into two categories: the analytical methods and the empirical methods. The analytical methods directly examine and assess the segmentation algorithms themselves by analyzing their principles and properties. The empirical methods indirectly judge the segmentation algorithms by applying them to test images and measuring the quality of segmentation results. Empirical methods are classified into two types: empirical goodness and empirical discrepancy method. In first method, properties of segmented images are measured using ”goodness” parameters. Where in the second type some references that present the ideal or expected segmentation results are first found.

Devices like Personal Data Assistants (PDA) which is pen input devices require good online handwriting character recognition algorithms. A. Funada et al. [18] proposed a new algorithm to recognize on-line handwriting and it utilize HMM (Hidden Markov Model). The memory reduction rate is a function of the matrix size and the number of states. They performed character segmentation, character classification which is fairly standard multilayer perceptron trained with error back propagation provides the ANN character classify

Table 1. Comparison of the different OCR Techniques

| Author(s) | Data set | Method | Recognition rate |
|---|---|---|---|
| Yaeger et al. [16] (1998) | (A-Z) characters, (0-9) digits, 23 symbols with writer independent system | Multi-Layer Perceptron | 21.3% |
| Hu et al. [17] (2000) | (a) 500, 1000 and 2000 unigen database. (b) 5000, 10000 and 20000 unigen database | Hidden Markov Model | 91.8%, 90.5% and 87.2% for (a) dataset and 83.2%, 79.8% and 76.3% for (b) dataset. |
| Funanda et al. [18] (2004) | Kanji, Katakana, Hiragana, Western alphabets and symbols with writer independent system | Hidden Markov Model | 91.34% |
| A. F. Mollah et al. [4] (2011) | Set of 100 business cards images | Segmentation using Vertical Projection Profile | 92.74% |
| M. Shen [7] (2015) | 1160 images with various resolutions, font sizes and noise levels. | Image Enhancement using non-linear Transformation | - |
| J.B. Pedersen et al. [3] (2016) | 100 images with a total of 640 characters | Character based segmentation and Nearest Neighbour Classifier | 98.69% |
| V. Kieu et al. [2] (2016) | IPAD contains 297 document images and PME contains 1998 document images | Fuzzy-C-Means clustering method | 90.57% |
| C. N. E. Anagnostopoulos et al. [21] (2006) | 1334 natural-scene gray-level vehicle images | probabilistic neural network (PNN) | 96.5% ( Segmentation) 89.1% (Entire Plate Recognition) |
| A. Coates et al. [8] | ICDAR data set 5198 test characters | Machine Learning Algorithm | 85.5% |

## 4. APPLICATION

Optical Character Recognition is a vast field with a number of varied application which is described below [23]. For OCR enhanced image segmentation algorithm based on histogram equalization using genetic algorithms are used.

### 4.1 Captcha

A CAPTCHA is a program that can generate and grade tests that human can pass but current computers programmers' cannot. In CAPTCHA, an image consisting of series of letters of number is generated which is obscured by image distortion techniques, size and font variation, distracting backgrounds, random segments, highlights, and noise in the image. This system can be used to remove this noise and segment the image to make the image tractable for the OCR (Optical Character Recognition) systems.

### 4.2 Institutional Repositories and Digital Libraries

Institutional repositories are digital collections of the outputs created within a university or research institution. It is an online locale of intellectual data of an institution, especially a research institution where it is collected, preserved and aired. It helps to open up the outputs of an institution and give it visibility and more impact on worldwide level

**4.3 Invoice Imaging** Invoice imaging is widely used in many businesses applications to keep track of financial records and prevent a backlog of payments from pilling up. In government agencies and independent organizations, OCR simplifies data collection and analysis, among other processes

**4.4 Automatic Number Recognition**

Automatic number plate recognition [6] is used as a mass surveillance technique making use of optical character recognition on images to identify vehicle registration plates. ANPR has also been made to store the images captured by the cameras including the numbers captured from the license plate.

**4.5 Legal Industry** The legal industry is also one of the beneficiaries of the OCR technology. OCR is used to digitize documents and directly entered into a computer database.

**4.6 Banking** Another important application of OCR is in banking, where it is used to process cheques without human involvement Cheque can be inserted into a machine where the system scans the amount to be issued and the correct amount of accessed as necessary

**4.7 Healthcare** Healthcare has also seen an increase in the use of OCR technology to process paperwork. Healthcare professionals always have to deal with large volumes of forms for each patient, including insurance forms as well as general health forms. To keep up with all of this information, it is useful to input relevant data into an electronic database that can be accessed as necessary

## 5. CONCLUSION

"This paper elaborated survey of disparate techniques for OCR" has been studied. Handwritten character, natural scene images, business cards and TV set images are selected for experimentation. A systematic flow of OCR system is discussed. In this paper projection profile based method for segmentation, fourier transform technique is for pre-processing, and nearest neighbour classifier for classification are described. This paper can be helpful to the researcher for selecting most appropriate techniques to achieve optimum results for application according to a different parameter described in the previous section

## 6. REFERENCES

[1] A. S. Sawant, "Script Independent Text Pre-processing and Segmentation for OCR," Int. Conf. Electr. Electron. Signals, Commun. Optim. - 2015, pp. 1–5, 2015.

[2] V. Kieu, F. Cloppet, and N. Vincent, "OCR Accuracy Prediction Method Based on Blur Estimation," 2016 12th IAPR Work. Doc. Anal. Syst., pp. 317–322, 2016.

[3] J. B. Pedersen, K. Nasrollahi, and T. B. Moeslund, "Quality Inspection of Printed Texts," IWSSP 2016- 23rd Int. Conf. Syst. Image Process. 23-25 May 2016, Bratislava, Slovakia, pp. 6–9, 2016.

[4] A. F. Mollah, N. Majumder, S. Basu, and M. Nasipuri, "Design of an Optical Character

Recognition System for Camera- based Handheld Devices," IJCSI, vol. 8, no. 4, pp. 283– 289, 2011.

[5] B. Jain and M. Borah, "A Comparison Paper on Skew Detection of Scanned Document Images Based on Horizontal and Vertical," IJSRP, vol. 4, no. 6, pp. 4–7, 2014.

[6] E. N. Bhatia, "Optical Character Recognition Techniques : A Review," IJARCSSE, vol. 4, no. 5, pp. 1219–1223, 2014.

[7] M. Shen, "Improving OCR Performance with Background Image Elimination," 2015 12th Int. Conf. Fuzzy Syst. Knowl. Discov., pp. 1566–1570, 2015.

[8] A. Coates et al., "Text Detection and Character Recognition in Scene Images with Unsupervised Feature Learning."

[9] P. Road, "Confidence Guided Progressive Search and Fast Match Techniques for High Performance ChineseEnglish OCR *," IEEE, pp. 89–92, 2002.

[10] H. Wang and J. Kangas, "Character-Like Region Verification for Extracting Text in Scene Images," no. 11, 2001.

[11] I. Kastelan, S. Kukolj, V. Pekovic, V. Marinkovic, and Z. Marceta, "Extraction of Text on TV Screen using Optical Character Recognition," IEEE, pp. 153–156, 2012.

[12] J. Diaz-escobar, "Optical Character Recognition based on phase features," IEEE, 2015.

[13] A. Thilagavathy, K. Aarthi, and A. Chilambuchelvan, "A Hybrid Approach to Extract Scene Text from Videos," ICCEET, pp. 1017–1022, 2012.

[14] S. Goyal, "Optical Character Recognition," IJARCSSE, vol. 3, no. 11, pp. 982–985, 2013.

[15] G. Vamvakas, B. Gatos, N. Stamatopoulos, and S. J. Perantonis, "A Complete Optical Character Recognition Methodology for Historical Documents," pp. 525–532, 2008.

[16] L. S. Yaeger, B. J. Webb, and R. F. Lyon, "Search for Online , Printed Handwriting N EWTON," Am. Assoc. Artif. Intell., vol. 19, no. 1, pp. 73–90, 1998.

[17] J. Hu, S. G. Lim, and M. K. Brown, "Writer independent on-line handwriting recognition using an HMM approach," J. PATTERN Recognit. Soc., vol. 33, pp. 133–147, 2000.

[18] A. Funada, D. Muramatsu, and T. Matsumoto, "The Reduction of Memory and the Improvement of Recognition Rate for HMM On-line Handwriting Recognition," IEEE, pp. 0–5, 2004.

[19] J. r´ı Matas, "Real-Time Scene Text Localization and Recognition," IEEE, pp. 3538–3545, 2012. [20] H. Lin and C. Hsu, "Optical

Character Recognition with Fast Training Neural Network," IEEE, pp. 1458–1461, 2016.

[20] C. N. E. Anagnostopoulos, I. E. Anagnostopoulos, V. Loumos, and E. Kayafas, "A License Plate-Recognition Algorithm for Intelligent Transportation System Applications," IEEE, vol. 7, no. 3, pp. 377–392, 2006.

[21] Y. J. Zhang, "A survey on evaluation methods for image segmentation," pp. 1–13.

[22] A. Singh, K. Bacchuwar, and A. Bhasin, "A Survey of OCR Applications," Int. J. Mach. Learn. Comput., vol. 2, no. 3, pp. 314–318, 2012.

# IMAGE BASED APPRAISAL OF REAL ESTATE PROPERTIES

**DEVELLA MOHITHA**
PG Scholar, Department of Computer Science,
B.V. RAJU College,
Bhimavaram, W.G.Dt., A.P, India

**V. BHASKAR MURTHY**
HOD & Associate Professor in Computer Science,
B.V. RAJU College,
Bhimavaram, W.G.Dt., A.P, India

**ABSTRACT**

Real estate appraisal, which is the process of estimating the price for real estate properties, is crucial for both buys and sellers as the basis for negotiation and transaction. Traditionally, the repeat sales model has been widely adopted to estimate real estate price. However, it depends the design and calculation of a complex economic related index, which is challenging to estimate accurately. Today, real estate brokers provide easy access to detailed online information on real estate properties to their clients. We are interested in estimating the real estate price from these large amounts of easily accessed data. In particular, we analyze the prediction power of online house pictures, which is one of the key factors for online users to make a potential visiting decision. The development of robust computer vision algorithms makes the analysis of visual content possible. In this work, we employ a Recurrent Neural Network (RNN) to predict real estate price using the state-of-the-art visual features. The experimental results indicate that our model outperforms several of other state-of-the-art baseline algorithms in terms of both mean absolute error (MAE) and mean absolute percentage error (MAPE).

**Index Terms—visual content analysis, real estate, deep neural networks**

## 1. INTRODUCTION

Real estate appraisal, which is the process of estimating the price for real estate properties, is crucial for both buys and sellers as the basis for negotiation and transaction. Real estate plays a vital role in all aspects of our contemporary society. In a report published by the European Public Real Estate Association (EPRA http://alturl.com/7snxx), it was shown that real estate in all its forms accounts for nearly 20% of the economic activity. Therefore, accurate prediction of real estate prices or the trends of real estate prices help governments and companies make informed decisions. On the other hand, for most of the working class, housing has been one of the largest expenses. A right decision on a house, which heavily depends on their judgment on the value of the property, can possibly help them save money or even make profits from their investment in their homes. From this perspective, real estate appraisal is also closely related to people's lives. Current research from both estate industry and academia has reached the conclusion that real estate value is closely related to property

infrastructure, traffic, online user reviews and so on. Generally speaking, there are several different types of appraisal values. In particular, we are interested in the market value, which refers to the trade price in a competitive Walrasian auction setting. Today, people are likely to trade through real estate brokers, who provide easy access online websites for browsing real estate property in an interactive and convenient way. Fig. 1 shows an example of house listing from Realtor (http://www.realtor.com/), which is the largest real estate broker in North America. From the figure, we see that a typical piece of listing on a real estate property will introduce the infrastructure data in text for the house along with some pictures of the house. Typically, a buyer will look at those pictures to obtain a general idea of the overall property in a selected area before making his next move. Traditionally, both real estate industry professionals and researchers have relied on a number of factors, such as economic index, house age, history trade and neighborhood environment [5] and so on to estimate the price. Indeed, these factors have been proved to be related to the house price, which is quite difficult to estimate and sensitive to many different human activities. Therefore, researchers have devoted much effort in building a robust house price index. In addition, quantitative features including Area, Year, Stores, Rooms and Centre are also employed to build neural network models for estimating house prices. However, pictures, which is probably the most important factor on a buyer's initial decision making process, have been ignored in this process. This is partially due to the fact that visual content is very difficult to interpret or quantify by computers compared with human beings. A picture is worth a thousand words. One advantage with images and videos is that they act like universal languages. People with different backgrounds can easily understand the main content of an image or video. In the real estate industry, pictures can easily tell people exactly how the house looks like, which is impossible to be described in many ways using language. For the given house pictures, people can easily have an overall feeling of the house, e.g. what is the overall construction style, how the neighboring environment looks like. These high-level attributes are difficult to be quantitatively described. On the other hand, today's computational infrastructure is also much cheaper and more powerful to make the analysis of computationally intensive visual content analysis feasible. Indeed, there are existing works on focusing the analysis of visual content for tasks such as prediction [13], [14], and online user profiling [15]. Due to the recently developed deep learning, computers have become smart enough to interpret visual content in a way similar to human beings. Recently, deep learning has enabled robust and accurate feature learning, which in turn produces the state-of-the-art performance on many computer vision related tasks, e.g. digit recognition, image

classification, aesthetics estimation and scene recognition. These systems suggest that deep learning is very effective in learning robust features in a supervised or unsupervised fashion. Even though deep neural networks may be trapped in local optima, using different optimization techniques, one can achieve the state-of-the-art performance on many challenging tasks mentioned above. Inspired by the recent successes of deep learning, in this work we are interested in solving the challenging real estate appraisal problem using deep visual features. In particular, for images related tasks, Convolutional Neural Network (CNN) are widely used due to the usage of convolutional layers. It takes into consideration the locations and neighbors of image pixels, which are important to capture useful features for visual tasks. Convolutional Neural Networks have been proved very powerful in solving computer vision related tasks.

We intend to employ the pictures for the task of real estate price estimation. We want to know whether visual features, which are a reflection of a real estate property, can help estimate the real estate price. Intuitively, if visual features can characterize a property in a way similar to human beings, we should be able to quantify the house features using those visual responses. Meanwhile, real estate properties are closely related to the neighborhood. In this work, we develop algorithms which only rely on 1) the neighbor information and 2) the attributes from pictures to estimate real estate property price. To preserve the local relation among properties we employ a novel approach, which employs random walks to generate house sequences. In building the random walk graph, only the locations of houses are utilized. In this way, the problem of real estate appraisal has been transformed into a sequence learning problem. Recurrent Neural Network (RNN) is particularly designed to solve sequence related problems. Recently, RNNs have been successfully applied to challenging tasks including machine translation image captioning [26], and speech recognition. Inspired by the success of RNN, we deploy RNN to learn regression models on the transformed problem. The main contributions of our work are as follows: To the best of our knowledge, we are the first to quantify the impact of visual content on real estate price estimation. We attribute the possibility of our work to the newly designed computer vision algorithms, in particular Convolutional Neural Networks (CNNs). We employ random walks to generate house sequences according to the locations of each house. In this way, we are able to transform the problem into a novel sequence prediction problem, which is able to preserve the relation among houses. We employ the novel Recurrent Neural Networks (RNNs) to predict real estate properties and achieve accurate results.

## 2. LITERATURE SURVEY

Literature survey is the most important step in software development process. Before developing the tool it is necessary to determine

the time factor, economy and company strength. Once these things are satisfied, ten next steps are to determine which operating system and language can be used for developing the tool. Once the programmers start building the tool the programmers need lot of external support. This support can be obtained from senior programmers, from book or from websites. Before building the system the above consideration are taken into account for developing the proposed system.

## 3. EXISTING SYSTEM

Current research from both estate industry and academia has reached the conclusion that real estate value is closely related to property infrastructure, traffic, online user reviews and so on. Generally speaking, there are several different types of appraisal values. In particular, we are interested in the market value, which refers to the trade price in a competitive walrasian auction setting Traditionally, both real estate industry professionals and researchers have relied on a number of factors, such as economic index, house age, history trade and neighborhood environment and so on to estimate the price. Indeed, these factors have been proved to be related to the house price, which is quite difficult to estimate and sensitive to many different human activities. The current algorithms are 1). Regression Models and 2). Deep Walk. Regression model has been employed to analyze real estate price index. Recently, the results in Fu et al. show that sparse regularization can obtain better performance in real estate ranking. Thus, we choose to use LASSO which is an l1-constrained regression model, as one of our baseline algorithms. Deep Walk is another way of employing random walks for unsupervised feature learning of graphs. The main approach is inspired by distributed word representation learning. In using DeepWalk, we also use _-neighborhood graph with the same settings with the graph we built for generating sequences for B-LSTM. The learned features are also fed into a LASSO model for learning the regressionweights. Indeed, deepwalk can be thought as a simpler versionof our algorithm, where only the graph structures are employedto learn features. Our framework can employ both the graph structure and other features, i.e. visual attributes, for buildingregression model.

### 3.1 DISADVANTAGE:

The existing system is quite difficult to estimate and sensitive to many different human activities. There are lot of difficult works have been done with the existing systems to measure the number of factors such as economic index, house age, history trade and neighborhood environment. Current research from both estate industry and academia has reached the conclusion that real estate value is closely related to property infrastructure, traffic online user Reviews and so on.

## 4. PROPOSED SYSTEM

We intend to employ the pictures for the task of real estate price estimation. We want to know whether visual features, which are a reflection of a real estate property, can help estimate the real estate price. Intuitively, if visual features can characterize a property in a way similar to human beings, we should be able to quantify the house features using those visual responses. Meanwhile, real estate properties are closely related to the neighborhood. In this work, we develop algorithms which only rely on

1) the neighbor information and

2) the attributes from pictures to estimate real estate property price To preserve the local relation among properties we employ a novel approach, which employs random walks to generate house sequences.

In building the random walk graph, only the locations of houses are utilized. In this way, the problem of real estate appraisal has been transformed into a sequence learning problem. Recurrent Neural Network (RNN) is particularly designed to solve sequence related problems. Recently, RNNs have been successfully applied to challenging tasks including machine translation, image captioning, and speech recognition. Inspired by the success of RNN, we deploy RNN to learn regression models on the transformed problem. The main contributions of our work are as follows: To the best of our knowledge, we are the first to quantify the impact of visual content on real estate price estimation. We attribute the possibility of our work to the newly designed computer vision algorithms, in particular Convolutional Neural Networks (CNNs). We employ random walks to generate house sequences according to the locations of each house. In this way, we are able to transform the problem into a novel sequence prediction problem, which is able to preserve the relation among houses. We employ the novel Recurrent Neural Networks (RNNs) to predict real estate properties and achieve accurate results.

### 4.1 ADVANTAGE:

A picture is worth a thousand words. One advantage with images and videos is that they act like universal languages. For the given house pictures, people can easily have an overall feeling of the house, e.g. what is the overall construction style, how the neighboring environment looks like. These high-level attributes are difficult to be quantitatively described. Map Based Location information are most commonly effective than the viewing in raw details. The most accurate details can be viewed in simple steps. The proposed algorithms are very effective than the existing algorithms such as LASSO and Deep Walk.

## 5. MODULES

### 5.1 PROPERTY ADDITION

The property addition is the main initiative module for the project. Once authorized user login into the system, they can perform their activity as per their wish. In this module, User

must have interested in selling the property which they own. The Property details such as Location, Address, and Facilities that the households are need to add to the cloud where everything that seller uploads can viewable to buyer and agent.

### 5.2 ADDING LOCATION DETAILS

In this module user that is seller need to upload the details of their location as well as their neighboring facility location such as schools, colleges and medical etc., In previous modules also user need to add the location that are into the raw typed format but here in this module we can upload the location details in maps and map formats. Spotting these locations can be very handy for agents or users to get to know about the details of property and neighboring details.

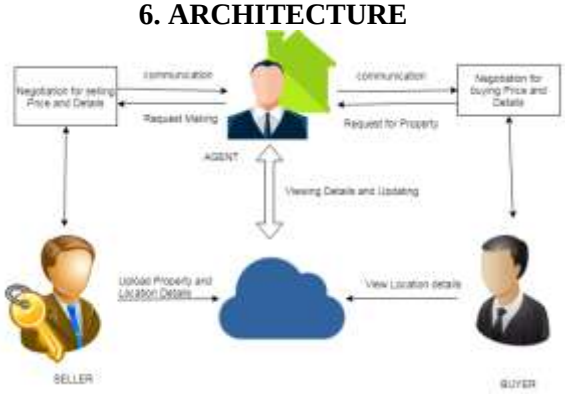
### 5.3 PRICE NEGOTIATION

This module is mainly designed for buyers and agents. Firstly, buyer sends the request to agents along with the cost of expectations and other query details about property. Once agents view the request from the buyer, Agent can decide the price according to the merit of location and both the buyer and seller. This module designed like chat. Dual way communication can be accomplished among the various users.

### 5.4 GEOMETRICAL ANALYSIS

The Geometrical analysis of given data set is done by charts. Here in this project there are two graphs have been plot between numbers of locations versus city. The pie chart and line charts are established in this project in order to analysis the data effectively.

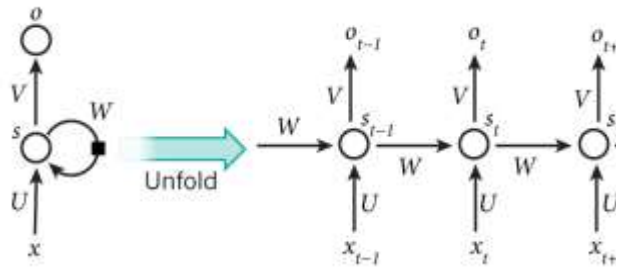
## 6. ARCHITECTURE




## 7. ALGORITHMS

### 7.1 Recurrent Neural Networks

The idea behind RNNs is to make use of sequential information. In a traditional neural network we assume that all inputs (and outputs) are independent of each other. But for many tasks that's a very bad idea. If you want to predict the next word in a sentence you better know which words came before it. RNNs are called *recurrent* because they perform the same task for every element of a sequence, with the output being depended on the previous computations. Another way to think about RNNs is that they have a "memory" which captures information about what has been calculated so far. In theory RNNs can make use of information in arbitrarily long sequences, but in practice they are limited to looking back only a few steps (more on this later). Here is what a typical RNN looks like:

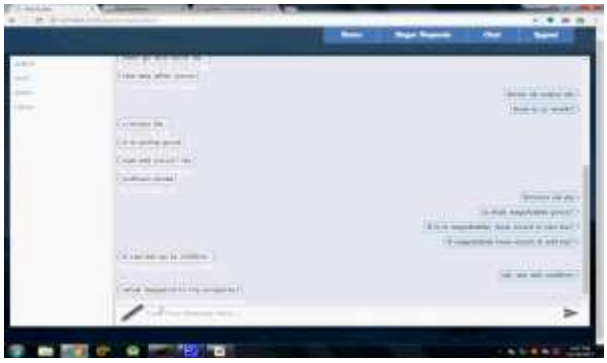


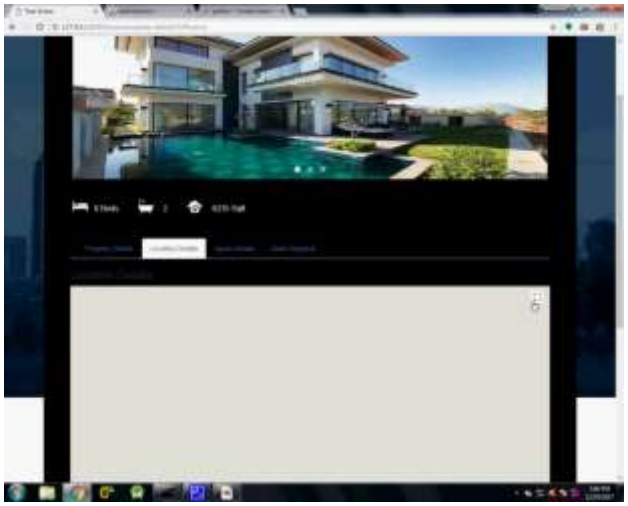

The above diagram shows a RNN being *unrolled* (or unfolded) into a full network. By unrolling we simply mean that we write out the network for the complete sequence. For example, if the sequence we care about is a sentence of 5 words, the network would be unrolled into a 5-layer neural network, one layer for each word.

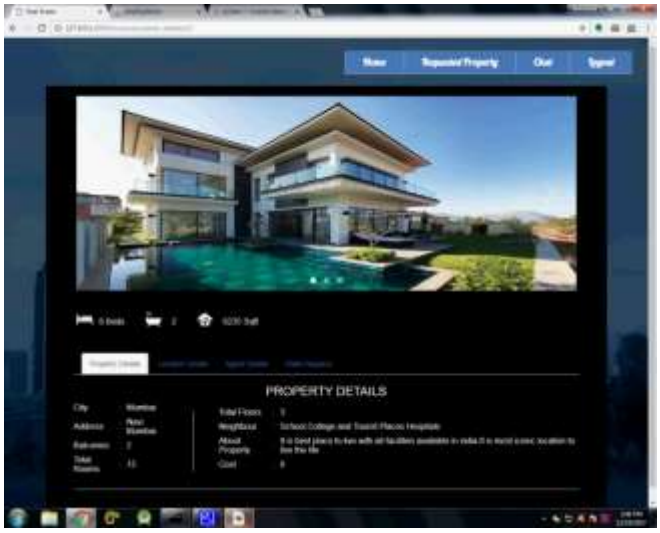
### 7.2 State Of The Art

**State of the art** (sometimes **cutting edge**) refers to the highest level of general development, as of a device, technique, or scientific field achieved at a particular time. It also refers to such a level of development reached at any particular time as a result of the common methodologies employed at the time.

## 8. OUTPUT RESULTS




PROPERTY DETAILS


MAKE REQUEST

## 9. CONCLUSION

In this work, we propose a novel framework for real estate appraisal. In particular, the proposed framework is able to take both the location and the visual attributes into consideration. The evaluation of the proposed model on two selected cities suggests the effectiveness and flexibility of the model. Indeed, our work has also offered new approaches of applying deep neural networks on graph structured data. We hope our model can not only give insights on real estate appraisal, but also can inspire others on employing deep neural networks on graph structured data.

## 10. REFERENCES

**1.**Y. Fu et al., "Exploiting geographic dependencies for real estate appraisal: A mutual perspective of ranking and clustering", *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, pp. 1047-1056, 2014.

**2.**K. Wardrip, "Public transits impact on housing costs: A review of the literature", *Center for Housing Policy*, Aug. 31 2011, [online] Available: http://www.reconnectingamerica.org/resource-center/browse-research/2011/public-transit-s-impact-on-housing- costs-a-review-of-the-literature/SearchForm/?Search=Keith+Wardrip.

**3.**Y. Fu et al., "Sparse real estate ranking with online user reviews and offline moving behaviors", *Proc. IEEE Int. Conf. Data Mining*, pp. 120-129, 2014.

**4.**A. Beja and M. B. Goldman, "On the dynamic behavior of prices in disequilibrium", *J. Finance*, vol. 35, no. 2, pp. 235-248, 1980.

**5.**E. L'Eplattenier, "How to run a comparative market analysis (CMA) the right way", 2016, [online] Available: http://fitsmallbusiness.com/comparative-market-analysis/.

**6.**M. J. Bailey, R. F. Muth and H. O. Nourse, "A regression method for real estate price index construction", *J. Amer. Statist. Assoc.*, vol. 58, no. 304, pp. 933-942, 1963.

**7.**R. Meese and N. Wallace, "Nonparametric estimation of dynamic hedonic price models and the construction of residential housing price indices", *Real Estate Econ.*, vol. 19, no. 3, pp. 308-332, 1991.

**8.**S. T. Anderson and S. E. West, "Open space residential property values and spatial context", *Region. Sci. Urban Econ.*, vol. 36, no. 6, pp. 773-789, 2006.

**9.**C. H. Nagaraja et al., "An autoregressive approach to house price modeling", *Annal. Appl. Statist.*, vol. 5, no. 1, pp. 124-149, 2011.

**10.**T. Lasota, Z. Telec, G. Trawiński and B. Trawiński, "Empirical comparison of resampling methods using genetic fuzzy systems for a regression problem" in Intelligent Data Engineering and Automated Learning-IDEAL, New York, NY, USA:Springer, pp. 17-24, 2011.

**11.**O. Kempa, T. Lasota, Z. Telec and B. Trawiński, "Investigation of bagging ensembles of genetic neural networks and fuzzy systems for real estate appraisal" in Intelligent Information

and Database Systems, New York, NY, USA:Springer, pp. 323-332, 2011.

**12.**W. Di, N. Sundaresan, R. Piramuthu and A. Bhardwaj, "Is a picture really worth a thousand words?:-On the role of images in e-commerce", *Proc. 7th ACM Int. Conf. Web Search Data Mining*, pp. 633-642, 2014.

**13.**X. Jin, A. Gallagher, L. Cao, J. Luo and J. Han, "The wisdom of social multimedia: Using flickr for prediction and forecast", *Proc. Int. Conf. Multimedia*, pp. 1235-1244, 2010.

**14.**Q. You, L. Cao, Y. Cong, X. Zhang and J. Luo, "A multifaceted approach to social multimedia-based prediction of elections", *IEEE Trans. Multimedia*, vol. 17, no. 12, pp. 2271-2280, Dec. 2015.

**15.**Q. You, S. Bhatia and J. Luo, "A picture tells a thousand words?About you! user interest profiling from user generated visual content", *Signal Process.*, vol. 124, pp. 45-53, 2016.

**16.**Y. LeCun et al., "Backpropagation applied to handwritten zip code recognition", *Neural Comput.*, vol. 1, no. 4, pp. 541-551, 1989.

**17.**G. E. Hinton, S. Osindero and Y.-W. Teh, "A fast learning algorithm for deep belief nets", *Neural Comput.*, vol. 18, no. 7, pp. 1527-1554, 2006.

**18.**D. C. Cireşan, U. Meier, J. Masci, L. M. Gambardella and J. Schmidhuber, "Flexible high performance convolutional neural networks for image classification", *Proc. 22nd Int. Joint Conf. Artif. Intell.*, pp. 1237-1242, 2011.

**19.**A. Krizhevsky, I. Sutskever and G. E. Hinton, "ImageNet classification with deep convolutional neural networks", *Proc. 25th Int. Conf. Neural Inf. Process. Syst.*, pp. 1097-1105, 2012.

**20.**X. Lu, Z. Lin, H. Jin, J. Yang and J. Z. Wang, "Rapid: Rating pictorial aesthetics using deep learning", *Proc. 22nd ACM Int. Conf. Multimedia*, pp. 457-466, 2014.

**21.**B. Zhou, A. Lapedriza, J. Xiao, A. Torralba and A. Oliva, "Learning deep features for scene recognition using places database", *Proc. 27th Int. Conf. Neural Inf. Process. Syst.*, pp. 487-495, 2014.

**22.**G. E. Hinton, "A practical guide to training restricted Boltzmann machines", 2010.

**23.**Y. Bengio, "Practical recommendations for gradient-based training of deep architectures" in Neural Networks: Tricks of the Trade, New York, NY, USA:Springer, pp. 437-478, 2012.

# Opinion Mining For Social Networking Sites

**KAGITA YUKTHA SAI PRIYA**
PG Scholar, Department of Computer Science,
B.V. RAJU College,
Bhimavaram, W.G.Dt., A.P, India

**S.K. ALISHA**
Senior Assistant Professor in Computer Science,
B.V. RAJU College,
Bhimavaram,W.G.Dt., A.P, India

**V. BHASKAR MURTHY**
HOD & Associate Professor in Computer Science,
B.V. RAJU College,
Bhimavaram, W.G.Dt., A.P, India

***Abstract-***
This system uses opinion mining methodology in order to achieve desired functionality. Opinion Mining for Social Networking Site is a web application. Here the user will post his views related to some subject other users will view this post and will comment on this post. The System takes comments of various users, based on the opinion, system will specify whether the posted topic is good, bad, or worst. User can change his own profile picture and can update his status. These changes can be viewed by various users. We use a database of sentiment based keywords along with positivity or negativity weight in database and then based on these sentiment keywords mined in user comment is ranked. Once the user logins to the system, user can view his own status as well as he can view the topics posted by other users. When the user clicks on a particular topic user can give his own comment about the topic. System will use database and will match the comment with the keywords in database and will rank the topic. User can edit his own profile and can change his profile picture. The role of the admin is to add post and adds keywords in database. This application can be used by users who like to post view about some events that is already held, or can post about the events that is going to be held. This application also works as an advertisement which makes many people aware about the topic posted. This system is also useful for the user's who need review about their new idea. This system is also useful for the user's who need review about any particular event that is posted.

***Index Terms: opinion mining, sentiment analysis, social computing, social networks***

## I. Introduction

Content growth in the Internet in recent years has made a huge volume of information available. This information is presented in different formats such as posts, news articles, comments, and reviews. Especially in the automotive, electronics and film sectors, customers have written reviews about products or their features. By collecting and analyzing these reviews, new customers find others' opinion about different features of the product. They can compare the products to each other to find the best one that meets their needs. Moreover, manufacturers will find out strengths and weaknesses of their products or those of their competitors.

In this way, manufacturers will solve the reported problems and use the business intelligence behind the analysis for future investments. From the sentiment perspective, there are two kinds of textual information, namely, facts and opinions. While facts are the objective statements about the nature of a product, opinions describe attitudes, appraisals, and emotions regarding a product, service, topic, or an issue. Although the majority of research focuses on building applications around facts, the recent trend in the area of text mining has been focused on building applications around opinions Sentiment analysis is an interdisciplinary field that crosses natural language processing, artificial intelligence, and text mining. Since most opinions are available in

the text format and its processing is easier than other formats, sentimentanalysis has emerged as a subfield of text mining [3]. It generally recognizes opinions of people expressed in text. The opinions could be judgments, evaluations, affective (or emotional) states, beliefs, or wishes. Sentiment analysis appeared in the literature in 1990 for the first time and then it became a major research topic in 2000. Classifying the polarity of a given text as positive or negative is the basic task of sentiment analysis. Due to its many aspects it is often referred to with different names such as opinion mining, sentiment classification, sentiment analysis, and sentiment extraction. It is widely believed that Sentiment analysis is needed and useful. It is also widely accepted that extracting sentiment from text is a hard semantic problem even for human beings. Additionally, sentiment analysis is domain specific, therefore the polarity of some terms depends on the context in which they are used. For example, while "small" for "size" as a feature in the electronic products is positive, in agricultural products such as fruit it has a negative polarity. Sentiment analysis is used in different domains such as shopping, entertainment, politics, education, marketing, and research and development. This paper focuses on sentiment classification in social domains.

From the technical perspective, two main approaches for sentiment analysis are Bag Of Words (BOW) and Feature Based Sentiment(FBS) [8]. In the BOW approach, each document is seen as a set of words. As a result, the syntactic and semantic information between words are lost. The BOW approach is not useful when opinions about products and their features have to be analyzed. In such cases, it is required to extract features. FBS has emerged as an approach for analyzing the sentiments of products and their features. The results of sentiment classification are presented in various formats in different domains: positive/negative, like/dislike, recommended/not-recommended, good/bad, buy / don't buy, excellent/boring(film), support/against [3], favorable/ unfavorable [10], bullish/bearish, or optimistic / pessimistic [1]

## II. Related Work

The literature on the analysis of social networks is extremely rich and extensive. The first suggestions for conducting social network analyzes came from the fields of social sciences and psychology [12] or economics [13]. Interestingly, much of this research reformulated what had previously been discussed in physics in the context of complex systems [14]. The most thorough summary of themes, models and algorithms of social network analysis can be found in [17]. Opinion mining is a relatively new domain that stretches between data mining, machine learning, and natural language processing. Sentiment analysis methods can be viewed as both supervised [1][5] and unsupervised learning methods [6][15] and as information retrieval methods [16][18]. Much work on opinion mining presents concepts based on handling text documents modeled as sets of words [1] or vectors, where dimensions represent words and values are weights of words in the document [2]. Semantic orientations of individual terms are aggregated using a dictionary method [5]. This method uses two small sets of manually identified positive and negative adjectives to serve as seed sets. New terms are subsequently added to these sentences if they are linked by semantically charged conjunctions such as and , but, but, etc. 6]. The semantic orientation of a term is inferred

from the assignment between the term and a word (or a set of words) that is clearly assigned to only one class (positive or negative),e.g. excellent and bad.

**N-GRAM FEATURES FOR AUTHOR PROFILING**

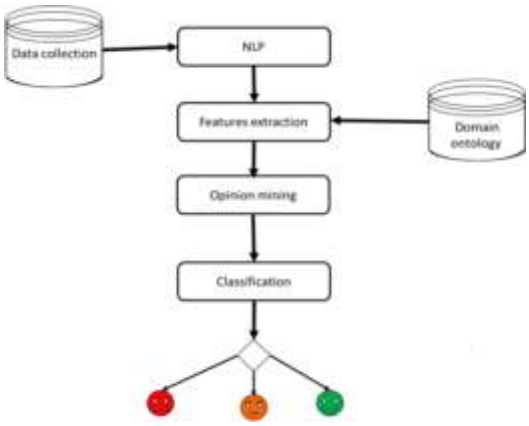
N-gram features can be classified into two categories: fixed and variable. Fixed n-grams are exact sequences occurring at either the character or token level. Variable n-grams are extraction patterns capable of representing more sophisticated linguistic phenomena. A plethora of fixed and variable n-grams have been used for opinion mining, including word, part-of-speech (POS), character, legomena, syntactic, and semantic n-grams. Word n-grams include bag-of-words (BOWs) and higher order word n-grams (e.g., bigrams, trigrams). Word n-grams have been used effectively in several studies [28]. Typically, unigrams to trigrams are used [3], [27], though 4- grams have also been employed [34]. Word n-grams often provide a feature set foundation, with additional feature categories added to them [4], [27], [34], [38]. Given the pervasiveness of adjectives and adverbs in opinion-rich text, POS tag, n-grams are very useful for sentiment classification [10], [12]. Additionally, some studies have employed word plus part-of-speech (POS Word) n-grams. These n-grams consider a word along with its POS tag in order to overcome word-sense disambiguation in situations where a word may otherwise have several senses [38]. For example, the phrase “quality of the” can be represented with the POS Word trigram “quality-noun of prep the-det.” Character n-grams are letter sequences. For example, the word “like” can be represented with the following two and three letter sequences “li, ik, ke, lik, ike.” While character n-grams were previously used mostly for style classification, they have recently been shown to be useful in related affect classification research attempting to identify emotions in text [2]. Legomena n-grams are collocations that replace once (hapax legomena) and twice occurring words (dis legomena) with “HAPAX” and “DIS” tags [2], [38].

**Lemmatization** is a process of identifying the lemma of a word. Algorithms to perform this operation typically use dictionaries in which they look up the primary form of the word. Lemmatization can find several different lemmas for a given word if the word is the inflected form of many different lemmas. Using lemmatization reduces the number of terms present in the corpus and allows matching of words in documents, even when words tend to appear in different grammatical forms. For text classification problems, however, the use of lemmatization can lead to a deterioration in the classification accuracy, since depending on which class the document belongs to, words can appear in different forms that are derived from a lemma.

**Stemming** is a similar process to lemmatization. It aims to extract the core of the word, called the stem, from the inflected word forms. Stemming typically involves removing and replacing prefixes and suffixes. The result of the stemming does not have to be, and often is not, a real lemma. The most well-known stemming algorithm is the Porter-Stemmer [7]. A stoplist is a set of words that should be removed at an early stage of word processing. In most cases, these are conjunctions and other words that do not add any additional information to the content of the sentence. Often stop-list words are present in the sentence only because of the requirements of the language grammar. In many cases, using stoplists improves the accuracy and performance of text document processing.

**A *term*** is a token generated from the document. It can be a word, lemma, stem, or n-gram. An n-gram is a sequence of n letters that appear in document content, e.g. the string Opinion Mining can be divided into the following 8 grams: Opinion_, Pinion_m, Inion_mi, Nion_Min, Ion_Mini, On_Minin, On_Minin, N_Mining. The n-gram representation of documents is often an alternative to term representation. N-grams are lossless because the text can be rebuilt, e.g. using DNA sequencing algorithms [8]. The n-gram representation enables the same operations on document collections as the term representation, but also offers extended functionality (e.g. spelling corrections).

**Architecture**




## III. EXISTING SYSTEM

- The concept of Sentimental analysis, motivated by different real-world applications and business-intelligence requirements, has recently been interpreted more broadly to include many different types of analysis of text, such as the treatment of opinion, sentiment or subjectivity Within this broad field, the most known problem is referred to as Sentiment Polarity classification, in which the problem of classifying documents by their overall sentiment is considered ,i.e. determining whether a review is positive or negative.

### Disadvantages

- In the polarity based approach the setting a polarity for a user review, which is a sentence is a difficult task because of the presence of sentences which have negative and positive information.
- Subjectivity and Objectivity identification sometimes difficult because of the presence of objective document in the subjectivity sentences.
- Sentiment lexican based methods to Author profiling here the sentiment lexicons are usually built without the supervision.

## IV. PROPOSED SYSTEM

The technology advancements now enable the different product developers to know about there products response in the market through the users reviews. They can maximize there profits and also gets a chance to improve their quality of product through Author profiling.

We propose the use of a rich set of n-gram features spanning many fixed and variable n-gram categories. We couple the extended feature set with a feature selection method capable of efficiently identifying an enhanced subset of n-grams for opinion classification. The proposed Feature Relation Network is a rule-based multivariate n-gram feature selection technique that efficiently removes redundant or less useful n-grams, allowing for more effective n-gram feature sets. Experimental results reveal that the extended feature set and proposed feature selection method can improve opinion classification performance over existing selection methods.

## Advantages

- ✓ The proposed feature selection method can improve opinion classification performance.
- ✓ The proposed Feature Relation Network is a rule-based multivariate n-gram feature selection technique that allowing for more effective n-gram feature sets.
- ✓ User can easily share his view about any subject.
- ✓ People can easily decide whether the posted topic is good or bad by using this application.
- ✓ This application is more useful for the users who love to comment.
- ✓ Since system ranks the topic based on the keywords in database so the result is appropriate.

## V.IMPLEMENTATION

- **Admin Login: -** Admin login's to the system using his Admin ID and password.
- **Add Post:** - Admin can post topics.
- **Add Keywords:** - Admin add keywords in database so that system will match the comment with the keywords in database and will rank the topic.
- **User Login: -** User login's to the system using his user ID and password.
- **Comment:** - User will post comment on the topic.
- **View Comment:** - User can view comment of other user's.
- **Rating Calculation:** - System will match the comment with the keywords in database and will rate the topic.
- **Edit Profile:** - User can edit his profile and can change his profile picture.
- **Status:-** User can view status and can change his status.

## VI.CONCLUSION

There are many interesting directions that can be explored in this project. We are interested in how Author profiling which can yield information regarding user's pulse about a product or service. This information will be very much helpful for the product developers or service providers. By implementing the Author profiling using machine learning approach creating a model which much accurate than any other implementation. We studying adverse opinions for given entities which are reviews and with machine learning algorithms we both train and test model, with review data by creating the model. We are also studying in analyzing the degree to which our sentiment indices predict future changes in popularity or market behavior.

## REFERENCES

[1] A. Abbasi and H. Chen, "CyberGate: A System and Design Framework for Text Analysis of Computer Mediated Communication," MIS Quarterly, vol. 32, no. 4, pp. 811-837, 2008.

[2] A. Abbasi, H. Chen, S. Thoms, and T. Fu, "Affect Analysis of Web Forums and Blogs Using Correlation Ensembles," IEEE Trans. Knowledge and Data Eng., vol. 20, no. 9, pp. 1168-1180, Sept. 2008.

[3] A. Abbasi, H. Chen, and A. Salem, "Author profiling in Multiple Languages: Feature Selection for Opinion Classification

in Web Forums," ACM Trans. Information Systems, vol. 26, no. 3, article no. 12, 2008.

[4] S. Argamon, C. Whitelaw, P. Chase, S.R. Hota, N. Garg, and S. Levitan, "Stylistic Text Classification Using Functional Lexical Features," J. Am. Soc. Information Science and Technology, vol. 58, no. 6, pp. 802-822, 2008.

[5] P.V. Balakrishnan, R. Gupta, and V.S. Jacobs, "Development of Hybrid Genetic Algorithms for Product Line Designs," IEEE Trans. Systems, Man, and Cybernetics, vol. 34, no. 1, pp. 468-483, Feb. 2004.

[6] A. Burgun and O. Bodenreider, "Comparing Terms, Concepts, and Semantic Classes in WordNet and the Unified Medical Language System," Proc. North Am. Assoc. Computational Linguistics Workshop, pp. 77-82, 2001.

[7] H. Cui, V. Mittal, and M. Datar, "Comparative Experiments on Sentiment Classification for Online Product Reviews," Proc. 21st AAAI Conf. Artificial Intelligence, pp. 1265-1270, 2006.

[8] S.R. Das and M.Y. Chen, "Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web," Management Science, vol. 53, no. 9, pp. 1375-1388, 2007.

[9] A. Esuli and F. Sebastiani, "SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining," Proc. Fifth Conf. Language Resources and Evaluation, pp. 417-422, 2006.

[10] Z. Fei, J. Liu, and G. Wu, "Sentiment Classification Using Phrase Patterns," Proc. Fourth IEEE Int'l Conf. Computer Information Technology, pp. 1147-1152, 2004.

[11] G. Forman, "An Extensive Empirical Study of Feature Selection Metrics for Text Classification," J. Machine Learning Research, vol. 3, pp. 1289-1305, 2004.

[12] M. Gamon, "Sentiment Classification on Customer Feedback Data: Noisy Data, Large Feature Vectors, and the Role of Linguistic Analysis," Proc. 20th Int'l Conf. Computational Linguistics, pp. 841- 847, 2004.

[13] M. Genereux and M. Santini, "Exploring the Use of Linguistic Features in Author profiling," Proc. Corpus Linguistics Conf., pp. 27-30, 2007.

[14] F. Fleuret, "Fast Binary Feature Selection with Conditional Mutual Information," J. Machine Learning Research, vol. 5, pp. 1531-1555, 2004.

[15] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene Selection for Cancer Classification Using Support Vector Machines," Machine Learning, vol. 46, pp. 389-422, 2002.

[16] I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," J. Machine Learning Research, vol. 3, pp. 1157-1182, 2003.

[17] M. Hall and L.A. Smith, "Feature Subset Selection: A Correlation Based Filter Approach," Proc. Fourth Int'l Conf. Neural Information Processing and Intelligent Information Systems, pp. 855-858, 1997.

[18] M. Hu and B. Liu, "Mining and Summarizing Customer Reviews," Proc. ACM SIGKDD, pp. 168-177, 2004.

[19] S. Kim and E. Hovy, "Determining the Sentiment of Opinions," Proc. 20th Int'l Conf. Computational Linguistics, pp. 1367-1373, 2004.

[20] J. Li, R. Zheng, and H. Chen, "From Fingerprint to Writeprint," Comm. ACM, vol. 49, no. 4, pp. 76-82, 2006.

[21] H. Liu and H. Motada, Feature Extraction, Construction, and Selection—Data Mining Perspective. Kluwer Academic Publishers, 1998.

[22] H. Liu and L. Yu, "Toward Integrating Feature Selection Algorithms for Classification and Clustering," IEEE Trans. Knowledge and Data Eng., vol. 17, no. 4, pp. 491-502, Apr. 2005.

[23] G. Mishne, "Experiments with Mood Classification," Proc. Stylistic Analysis of Text for Information Access Workshop, 2005.

[24] D. Mladenic, J. Brank, M. Grobelnik, and N. Milic-Frayling, "Feature Selection Using Linear Classifier Weights: Interaction with Classification Models," Proc. ACM SIGIR, pp. 234-241, 2004.

[25] S. Morinaga, K. Yamanishi, K. Tateishi, and T. Fukushima, "Mining Product Reputations on the Web," Proc. ACM SIGKDD, pp. 341-349, 2002.

[26] T. Nasukawa and T. Nagano, "Text Analysis and Knowledge Mining System," IBM Systems J., vol. 40, no. 4, pp. 967-984, 2001.

[27] V. Ng, S. Dasgupta, and S.M.N. Arifin, "Examining the Role of Linguistic Knowledge Sources in the Automatic Identification and Classification of Reviews," Proc. Conf. Computational Linguistics, Assoc. for Computational Linguistics, pp. 611-618, 2006.

[28] B. Pang, L. Lee, and S. Vaithyanathain, "Thumbs Up? Sentiment Classification Using Machine Learning Techniques," Proc. Conf. Empirical Methods in Natural Language Processing, pp. 79-86, 2002.

[29] B. Pang and L. Lee, "A Sentimental Education: Sentimental Analysis Using Subjectivity Summarization Based on Minimum Cuts," Proc. 42nd Ann. Meeting of

the Assoc. Computational Linguistics, pp. 271-278, 2004.

[30] F. Peng, D. Schuurmans, V. Keselj, and S. Wang, "Automated Authorship Attribution with Character Level Language Models," Proc. 10th Conf. European Chapter of the Assoc. Computational Linguistics, 2003.

[31] A. Popescu and O. Etzioni, "Extracting Product Features and Opinions from Reviews," Proc. Human Language Technology, Empirical Methods in Natural Language Processing, pp. 339-346, 2005.

[32] E. Riloff and J. Wiebe, "Learning Extraction Patterns for Subjective Expressions," Proc. Conf. Empirical Methods in Natural Language Processing, pp. 105-112, 2003.

[33] E. Riloff, J. Wiebe, and T. Wilson, "Learning Subjective Nouns Using Extraction Pattern Bootstrapping," Proc. Seventh Conf. Natural Language Learning, pp. 25-32, 2003.

[34] E. Riloff, S. Patwardhan, and J. Wiebe, "Feature Subsumption for Opinion Analysis," Proc. Conf. Empirical Methods in Natural Language Processing, pp. 440-448, 2006.

[35] K. Tsutsumi, K. Shimada, and T. Endo, "Movie Review Classification Based on Multiple Classifier," Proc. 21st Pacific Asia Conf. Language, Information, and Computation, pp. 481-488, 2007.

[36] P.D. Turney and M.L. Littman, "Measuring Praise and Criticism: Inference of Semantic Orientation from Association," ACM Trans. Information Systems, vol. 21, no. 4, pp. 315-346, 2003.

# HAND GESTURE RECOGNITION SYSTEM

**Kondeti Harshaditya**
PG Scholar, Department of Computer Science,
B.V. RAJU College,
Bhimavaram, W.G.Dt., A.P, India

**S.K.Alisha**
Senior Assistant professor in Computer Science,
B.V. RAJU College,
Bhimavaram,W.G.Dt., A.P, India

**V.Bhaskar Murthy**
HOD & Associate Professor in Computer Science,
B.V. RAJU College,
Bhimavaram, W.G.Dt., A.P, India

***Abstract-*** Hand gesture recognition (HGR) provides an intelligent and natural way of human computer interaction (HCI). Its applications range from medical rehabilitation to consumer electronics control (e.g. mobile phone). In order to distinguish hand gestures, various kinds of sensing techniques are utilized to obtain signals for pattern recognition. The HGR system can be divided into three parts according to its processing Steps : hand detection, finger identification, and gesture recognition. The system has two major advantages. First, it is highly modularized, and each of the three steps is capsuled from others; second, the edge/contour detection of hand as well as gesture recognition is an add-on layer, which can be easily transplanted to other applications. In IP Gaming we are proposing a system in which without using sensors and devices, we are detecting the hand and gesture with simple web camera and performing the image processing technique in which using those gesture, we can play game on console. In Image Process Gaming, the motions are detected through a web camera. These images are then passed for the image processing. The techniques used for image processing are hand gesture detection, edge detection, thresholding, contour detection. Using OpenCV, which provides a library collection of functions for different image processing techniques, these input images can be processed and corresponding key strokes will be generated.

***Index Terms*** Gesture recognition, Feature extraction, Assistive technology, Image segmentation, Algorithm design and analysis, Lighting, Image color analysis.

## I. Introduction

Hand gesture recognition provides an intelligent and natural way of human computer interaction (HCI). Its applications range from medical rehabilitation to consumer electronics control (e.g. mobile phone). In order to distinguish hand gestures, various kinds of sensing techniques are utilized to obtain signals for pattern recognition. Acceleration-base and electromyogram-based techniques are two research branches in the field of hand gesture pattern recognition. Acceleration-based (ACC-based) gesture control is usually studied as a supplementary interaction modality. It is well suited to distinguish noticeable, larger scale gestures with different hand trajectories of forearm movements. With ACC-based techniques some subtle finger or hand movement may be ignored whereas electromyogram-based (EMG-based) gesture recognition techniques use multi-channel EMG signals which contain rich information about hand gestures of various size scales. Due to some problems inherent in the EMG measurements, including the separability and reproducibility of measurement, the size of discriminable hand gesture set is still limited to 4-8 classes. In order to realize a natural and robust gesture-based HCI system, the selection of input hand gestures that are well discriminable from each other is of crucial

importance. Considering the complementary features of ACC- and EMG-measurements, we believe that their combination will increase the number of discriminable hand, wrist and forearm gestures and the accuracy of the recognition system. This paper describes In IP Gaming we are proposing a system in which without using sensors and Devices, we are detecting the hand and gesture with Simple Web camera and performing the Image Processing technique in which using those gesture, we can play game on console. In Image Process Gaming, the motions are detected through a web camera. These images are then passed for the image processing. The techniques used for image processing are hand gesture detection, edge detection, thresholding, contour detection. Using OpenCV, which provides a library collection of functions for different image processing techniques, these input images can be processed and corresponding key strokes will be generated.

## II.RELATED WORK

Motion capture and depth sensing are two emerging areas of research in recent years. With he launch of Kinect in 2010, Microsoft opened doors for researchers to develop, test and optimize the algorithms for these two areas. J Shotton proposed a method to quickly and accurately predict 3D positions of the body joints without using any temporal data. Key prospect of the method is they are considering a single depth image and are using a object recognition approach. From a single input depth image, they inferred a per pixel body part distribution. Leyv and T discussed about the Kinect technology. His work throws light on how the Identity of a person is tracked by the Kinect for XBox 360 sensor. Also a bit of information about how the changes are happening in the technology over the time is presented. With the launch of Kinect, there is a sea change in the identification and tracking techniques. The authors discussed the possible challenges over the next few years in the domain of gaming and Kinect sensor identification and tracking. Kinect identification is done by two ways: Biometric sign-in and Session tracking. A method to track fingertips and the centers of palms using Kinect was presented by Raheja. It applied thresholding on the depth of hand regions for segmentation. Then the palm was filtered and subtracted from the hand, so that only the fingers were left in the image. Under most situations when the hand was in front of the user, the fingers should be closest to the Kinect with the shallowest depth. Therefore, by determining the minimum depth, fingertips were found. The center of the palm was determined by finding the maximum of distance within the image of the hand. When fingers were extended, the accuracy of detecting fingertips was nearly 100% accuracy, and that of palm centers was around 90%. However this method did not attempt at gesture recognition. He proposed another approach using depth data provided by Kinect to detect fingertips. First, it found hand points by thresholding on depth elata, and then generated the convex hull containing the hand by Graham Scan. Fingertips were detected by calculating the angle between candidate points. After fingertips were found, the mouse clicking motion was recognized and tested on the popular game Angry Bird; that is, it recognized only one gesture.

## III. SYSTEM DESIN

- **PRE-PROCESSING**

Like many other pattern recognition tasks, pre-processing is necessary for enhancing robustness and recognition accuracy. The pre-processing prepares the image sequence for the recognition, so before calculating the diagonal Sum and other algorithms, a pre-processing step is performed to get the appropriate image, which is required for real time classification. So, it consists of some steps. The net effect of this processing is to extract the hand only from the given input because once the hand is detected from the given input it can be recognized easily .So, pre processing step mainly consists of following tasks:

- **SKINMODELLING**

There are numerous method used for skin detection such as RGB (Red, Green, Blue), YCbCr (Luminance Chrominance) and HSV (Hue, Saturation, Value).

- **RGB**:

RGB is a 3D color space pixel where each pixel has combination of three colors Red, Green and Blue at specific location. This technique widely used in image processing for identifying skin region

- **HSV (Hue,SaturationandValue):**

In HSV, Hue detect dominant color and Saturation define colorfulness whilst Value measure intensity or brightness. This is well enough to choose single color but it ignores complexity of color appearance. It trade off computation speed mean computationally expensive and perceptual relevance.

- ***Skin Detection:***

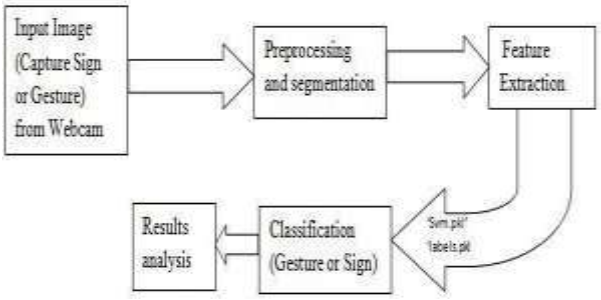
The skin color detection is one of important goal in hand gesture recognition. Skin color detection decision rules which we have to build that will discriminate between skin portion and non-skin portion pixels. This is accomplished usually by metric introduction ,which measure distance of the pixel color. This metric type is knowing as skin modelling.

- ***Removalo*** Background:

I have found that background greatly affects the result sofhand detection that's why I have decided to remove it. For this I have written our own code in spite of using any built-in ones All algorithms accept an input in RGB form and then convert it into binary format in order to provide ease in recognizing ny gesture and also retaining the luminance factor in an image.

- **Hand detection:**

Image could have more than one skin area but we required only hand for further process. For this I choose criteria image labeling which is following:




## IV. PROPOSED METHOD

Problem of gesture recognition in real time that sign language used by the community of deaf people. Research problem identified is based on Digital Image Processing using Colour Segmentation, Skin Detection, Image Segmentation, Image Filtering, and Template Matching techniques. This system recognizes gestures of including the alphabet and a subset of its words. Our system contains Data collection,

Data preparation, Image analysis, Feature extraction, Image classification, Image to text conversion.

**Construction of Convolutional Neural Network (CNN)**

The CNN is built using 6 different layers. The first layer is a Sequential layer which is used as a linear stack, where the instance of sequential layer is created and then the further layers will be added to it sequentially. The combination of Convolution and Pooling layer will continuously extract the features by down sampling the input images. The obtain results are fed into Flatten layer, where the input image is flattened which give rise to N-dimensional vector,where N is the number of classes from which the model selects the desired class. Once flattening of input image is completed the Dropout layer is used to prevent the model from overfitting by making the weights of some redundant neurons in a particular layer equal to zero. The Activation function such as Relu layer and Softmax layer are used to classify the images respectively to their classes.

**EDGE DETECTION**

Edge detection is one of the most commonly used operations in image analysis, and there are probably more algorithms in the literature for enhancing and detecting edges than any other single subject. The reason for this is that edges form the outline of an object. An edge is the boundary between an object and the background, and indicates the boundary between overlapping objects. This means that if the edges in an image can be identified accurately, all of the objects can be located and basic properties such as area, perimeter, and shape can be measured. Since computer vision involves the identification and classification of objects in an image, edge detections is an essential tool.

**CANNY**

The most significant new dimension to the canny algorithm is that it tries to assemble the individual edge candidate pixels into contours. These contours are formed by applying a hysteresis threshold to the pixels. This means that there are two thresholds, an upper and a lower. If a pixel has a gradient larger than the upper threshold, then it is accepted as an edge pixel; if a pixel is below the lower threshold, it is rejected. If the pixel's gradient is between the thresholds, then it will be accepted only if it is connected to a pixel that is above the high threshold.

**V. CONCLUSION**

In this paper, we have discussed how using camera can be used for Detection hand gestures and can be applied to any game control. We are using camera as a detecting device as well as input device for Augmented Reality System. The proposed system helps reduce the burden on experts to look into few regular activities. Instead, they can use our system for such activities. Also, the work simplifies the documentation process. The supervisor can keep track of current status of activity from his desk. Also, stepwise verification is possible as the system keeps track of each step. Through the introduction of our system, we will bring new opportunities for mechanical engineering based companies to use Augmented Reality for simplification of their complex tasks. This will add new dimensions to the conventional way of maintenance activities.

**REFERENCES**

[1] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore,A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. In Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, pages 1297 –1304, June 2011.

[2] T. Leyvand, C. Meekhof, Y.-C. Wei, J. Sun, and B. Guo. Kinect identity: Technology and experience. Computer, 44(4):94 –96, April 2011.

[3] S. Henderson and S. Feiner. Exploring the benefits of augmented reality documentation for maintenance and repair. Visualization and Computer Graphics, IEEE Transactions on, 17(10):1355 –1368, Oct. 2011.

[4] S. Henderson and S. Feiner. Augmented reality for maintenance and repair (armar). Technical report, AFRL-RHWP-TR-2007-0112, United States Air Force Research Lab,, Jul 2007.

[5] Saket Warade, Jagannath Aghav, Petitpierre Claude, Sandeep Udayagiri. Real-Time Detection and Tracking with Kinect. International Conference on Computer and Information Technology (ICCIT'2012) June 16-17, 2012, Bangkok

[6] R. Owens, "Lecture 6", Computer Vision IT412, 10/29/1997. http://homepages.inf.ed.ac.uk/rbf/CVonline/LOCAL_COPIES/ OWENS/LECT6/node2.html [7] S. Price, "Edges: The Canny Edge Detector", July 4, 1996. http://homepages.inf.ed.ac.uk/rbf/CVonline/LOCAL_COPIES/ MARBLE/low/edges/canny.html

[8] Ehsan Nadernejad, Sara Sharifzadeh, Hamid Hassanpour. Edge Detection Techniques: Evaluations and Comparisons. Applied Mathematical Sciences, Vol. 2, 2008, no. 31, 1507 – 1520.

# CoDetect Financial Fraud Detection with Anomaly Feature Detection

**K. VAMSI MAHESH**
PG Scholar, Department of Computer Science,
B.V. RAJU College,
Bhimavaram, W.G.Dt., A.P, India

**S.K. ALISHA**
Senior Assistant professor in Computer Science,
B.V. RAJU College,
Bhimavaram,W.G.Dt., A.P, India

**V. BHASKAR MURTHY**
HOD & Associate Professor in Computer Science,
B.V. RAJU College,
Bhimavaram, W.G.Dt., A.P, India

## *Abstract-*

Financial fraud, such as money laundering, is known to be a serious process of crime that makes illegitimately obtained funds go to terrorism or other criminal activity. This kind of illegal activities involve complex networks of trade and financial transactions, which makes it difficult to detect the fraud entities and discover the features of fraud. Fortunately, trading/transaction network and features of entities in the network can be constructed from the complex networks of the trade and financial transactions. The trading/transaction network reveals the interaction between entities, and thus anomaly detection on trading networks can reveal the entities involved in the fraud activity; while features of entities are the description of entities, and anomaly detection on features can reeffect details of the fraud activities. Thus, network and features provide complementary information for fraud detection, which has potential to improve fraud detection performance. However, the majority of existing methods focus on networks or features information separately, which does not utilize both information. In this paper, we propose a novel fraud detection framework, CoDetect, which can leverage both network information and feature information for financial fraud detection. In addition, the CoDetect can simultaneously detecting financial fraud activities and the feature patterns associated with the fraud activities. Extensive experiments on both synthetic data and real-world data demonstrate the efficiency and the effectiveness of the proposed framework in combating financial fraud, especially for money laundering.

***Index Term***
Gesture recognition, Feature extraction, Assistive technology, Image segmentation, Algorithm design and analysis, Lighting, Image color analysis.

## I. Introduction

In recent years, financial fraud activities such as credit card fraud, money laundering, increase gradually. These activities use the loss of personal and/or enterprises' properties. Even worse, they endanger the security of nation because the port from fraud may go to terrorism [1], [25]. Thus, accurately detecting financial fraud and tracing fraud are necessary and urgent. However, financial fraud detection is not an easy task due to the complex trading networks and transactions involved. Taking money laundering as an example, money laundering is denied as the process of using trades to move money/goods with the intent of obscuring the true origin of funds. Usually, the prices, quantity or quality of good son an invoice of money laundering are fake purposely. The misrepresentation of prices, quantity or quality of goods on an invoice merely exposes slight difference from regular basis if we use these numbers as features to generate detection policy. Under certain circumstances, this kind of detector may work well with relatively stable trading entities. Unfortunately, the real world situation is more

complicated, especially within Free Trade Zones (FTZs) where international trade involves complex procedures and exchange of information between trading entities. The fraud activities, especial money laundering, are deeper stealth. Money laundering activities may take different forms [1] such as the concealing transportation of cash using trading operations; the acquisition and sale of intangibles; and related party transactions. Not only the trading of goods shows on much more diversity, but also different type of companies, shell and front companies involve in to facilitate money laundering. In contrast with other fraud activities, money laundering demonstrates special characteristic which presents high risk to financial system with obscuring the money trail, collectivization behavior and wild trading regions in FTZs. Many fraud detection models work with attribute value data points that are generated from transactions data. Some aggregation methods are also used to enrich the information of data [28]. After generating feature points from transactions, supervised and unsupervised methods can be used to perform detection [26], [27], [34]. Usually, these data points are assumed to be independent and identically distributed(i.i.d.). However, the characteristic of money laundering is different from attribute-value data. The collectivization behavior means the data is inherently linked or partly linked. Obviously, trading activity involves at least two business entities. Linked data is patently not independent and identically distributed, which contradicts the assumptions of traditional supervised and unsupervised methods. On the other side, some linked data is auto correlated. For example, trading between business entity A and B implies that feature points A and B are correlated. Some features used to describe the properties of trading goods can be identical between A and B.This characteristic of auto correlation reduce the effective size of data for learning. Furthermore, feature points don't fuse the interaction information in data. The relations between any business entities indicate the potential causality that means, if businesses on going, fraud entity can be located by other identified fraud entity. This means the entity, which have connection with fraud entity, are suspicious. Consequently, feature based detection models with supervised or unsupervised methods have inherent limitation of incapacity of identifying what the fraud relations are. what the fraud relations are. Graph-based mining methods are one of the most important theories that attempt to identify relations between data points [3], [7], [13], as Fig. 1(a) shows. Financial activities can be modeled as a directed graph, then a sparse adjacent matrix can represent this graph. With graph-mining method, the sparse matrix can be approximated as summation of low-rank matrix and outlier matrix. The outlier matrix is a sign of suspicious fraud activities. Exploiting the graph based mining provides a new perspective for fraud detection and enables us to do advanced research on fraud detection. With the fraud activities detected by graph-based detection technique we are able to draw the conclusion that several business entities involved in fraud, however, we still don't know how these fraud activities are operated and why these activities labeled as fraud, i.e., the detailed features of the fraud activities. The majority of this how-and-why information is fused in features points, which have essential meaning for financial fraud detection because of the tracing necessity. For example, doing business with misrepresentation of the price may transfer additional value to exporter. The value in this example reveals how the fraud happened. This simple example requires the detection system to mark value as fraud property. Another example, fraud activities might go deeper stealth with multi-entities involved. If the same good or service invoices a number of different business entities to make the

payments, then there are several properties should be consider as suspicious: business location, name, direction, good or service etc. With the knowledge of the sesuspicious properties, tracing fraud can be much easier for executives.

## II. RELATED WORK

**Chang** proposed that situation analytics can be included in software service requirement analysis, which can facilitate the analysis of any change in user's requirements. Such an analysis is useful to understand the dynamic needs of a software service environment.

**Zhang *et al.* presented** a framework to the discovery of user behavior pattern in multimedia video recommendation services on online social networks. Their framework is based on social context and analyzes the changes in user need for different social situations. Such user behavior data can be obtained if we have access to user's logs] or user's clickstreams (e.g., recorded by social network platforms). The difference in user behavior can be obtained, for example, by analyzing the image search logs of users to study the search intention of different users , and this approach can facilitate optimization of search engines.

**Wang *et al.* [16]** used user clickstream data to construct a clickstream graph model to represent user behavior and identify different user groups, in order to detect malicious accounts. There have also been other researches that indicate user intent and abnormal accounts can be determined through behavior analysis, and social situation in facilitating the understanding of users' dynamic behavior.

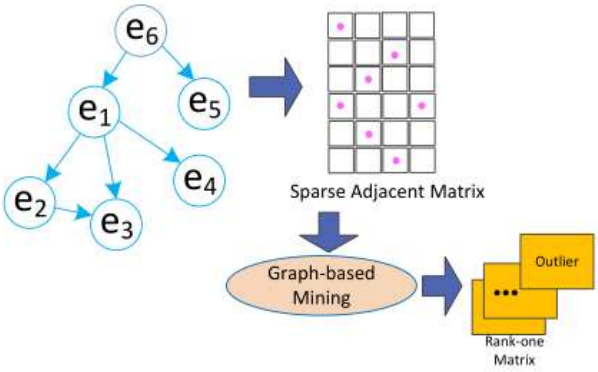
**Liu *et al.* [17]** constructed a new convolutional neural network architecture based on user behavior, search engine contentand context information to construct a click model and _nd out the user's click preferences to improve search quality.

**Al-Qurishi *et al.* [18]** collected a large amount of user information. on the Twitter and YouTube, about 13 million channel activities, analyzing and detecting abnormal behaviors that deviate significantly from large-scale specifications through user behavior in two social networks.

## III. EXISTING METHOD

Many fraud detection models work with attribute-valued ata points that are generated from transactions data. Some aggregation methods are also used to enrich the information of data [28]. After generating feature points from transactions, supervised and unsupervised methods can be used to perform detection [26], [27], [34]. Usually, these data points are assumed to be independent and identically distributed (i.i.d.). However, the characteristic of money laundering is different from attribute-value data. The collectivization behavior means the data is inherently linked or partly linked. Obviously, trading activity involves at least two business entities. Linked data is patently not independent and identically distributed, which contradicts the assumptions of traditional supervised and unsupervised methods. On the other side, some linked data is auto correlated. For example, trading between business entity A and B implies that feature points A and B are correlated. Some features used to describe the properties of trading goods can be identical between A and B. This characteristic of auto correlation reduces the effective size of data for learning. Furthermore, feature points don't fuse the interaction information in data. The relations between any business entities indicate the potential causality that means, if businesses on going, fraud entity can be located by other

identi_ed fraud entity. This means the entity, which have connection with fraud entity, are suspicious. Consequently, feature based detection models with supervised or unsupervised methods have inherent limitation of incapacity of identifying what the fraud relations are.




**Fraud detection using graph mining techniques**.

**Graph-based mining** methods are one of the most important theories that attempt to identify relations between data points [3], [7], [13], as Fig. 1(a) shows. Financial activities can be modeled as a directed graph, then a sparse adjacent matrix can represent this graph. With graph-mining method, the sparse matrix can be approximated as summation of low-rank matrix and outlier matrix. The outlier matrix is a sign of suspicious fraud activities. Exploiting the graph based mining provides a new perspective for fraud detection and enables us to do advanced research on fraud detection. With the fraud activities detected by graph-based detection technique we are able to draw the conclusion that several business entities involved in fraud, however, we still don't know how these fraud activities are operated and why these activities labeled as fraud, i.e., the detailed features of the fraud activities. The majority of this how-and-why information is fused in features points, which have essential meaning for _nancial fraud detection because of the tracing necessity. For example, doing business with misrepresentation of the price may transfer additional value to exporter. The value in this example reveals how did the fraud happen. This simple example requires the detection system to mark value as fraud property. Another example, fraud activities might go deeper stealth with multi-entities involved. If the same good or service invoices a number of different business entities to make the payments, then there are several properties should be consider as suspicious: business location, name, direction, good or service etc. With the knowledge of these suspicious properties, tracing fraud can be much easier for executives.

Thus, graph-based methods can detection suspicious interactions between entities while attribute-feature based methods can reveal the features of the fraud. Graph and attributes provides two complementary information for financial fraud activity detection and fraud property tracing. However, the majority of the existing algorithms exploits these two information separately and thus can not provide a system that can detect the fraud entities and reveal suspicious properties for easy tracing simultaneously.

## IV. PROPOSED METHOD

We would like to develop a novel framework for fraud detection by considering the special detecting and tracing demanding of fraud entities and behaviors. Specially, we investigate:

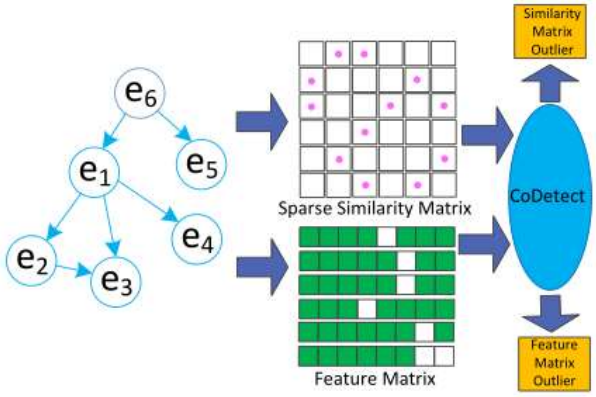
(1) How to utilize both graph matrix and feature matrix for fraud detection and fraud tracing;

(2) How to mathematically model both graph matrix and feature matrix so as to simultaneously achieve the tasks of fraud detection and tracing. In an attempt to solve these challenges, we proposed a novel detection framework CoDetect, for financial data, especially for money laundering data. We incorporate fraud entities detection and

anomaly feature detection in the same framework to and fraud patterns and corresponding features simultaneously. Combining entities detection and feature detection enables us to build a novel fraud detection framework for noisy and sparse financial data: relevant fraud patterns help the identification of fraud identities, and relevant features in turn help revealing of the nature of fraud activities.

Our empirical study on synthetic and real world data sets demonstrates the effectiveness of CoDetect, which does discover the fraud pattern and decide the fraud related properties in an unsupervised manner by seeking the low-rank approximation representations and residual for complex network matrix and feature matrix simultaneously. The major contributions of the paper can be summarized as follows:

1) Provide an approach to establish weighted graph from financial network, incorporating properties of nodes and links;
2) Demonstrate different scenarios of financial fraud and formulate the patterns of fraud in term of graph and sparse matrix;
3) Propose a novel unsupervised framework, CoDetect, for the problem of complex patterns discovery and anomaly features identification, employing two matrices residual analysis on graph-based financial network;
4) Evaluate framework using synthetic and real world data to demonstrate both effectiveness and efficiency of the proposed framework.




**The Proposed Framework.**

## V. CONCLUSION

We propose a new framework, CoDetect, which can perform fraud detection on graph-based similarity matrix and feature matrix simultaneously. It introduces a new way to reveal the nature of financial activities from fraud patterns to suspicious property. Furthermore, the framework provides a more interpretable way to identify the fraud on sparse matrix. Experiment a l results on synthetic and real world data sets show that the proposed framework (CoDetect) can effectively detect the fraud patterns as well as suspicious features. With this co detection framework, executives in financial supervision cannot only detect the fraud patterns but also trace the original of fraud with suspicious feature. Financial activities are involving with time. We can represent these activities into similarity tensor and feature tensor. So we would like to study how to integrate tensor into codetect framework for fraud detection.

**REFERENCES**

[1] C. Sullivan and E. Smith. ``Trade-Based Money Laundering: Risks andRegulatory Responses," Social Sci. Electron. Publishing, 2012, p. 6.
[2] United Press International. (May 2009). *Trade-Based MoneyLaundering Flourishing*.[Online].Available:http://www.upi.com/TopNews/2009/05/11/Trade-based-money-laundering-_ourishing/UPI-17331242061466
[3] L. Akoglu, M. McGlohon, and C. Faloutsos, ``OddBall: Spotting anomaliesin weighted graphs," in *Proc. Pacic-Asia Conf. Knowl. Discovery DataMining*, 2010, pp. 410_421.
[4] V. Chandola, A. Banerjee, and V. Kumar, ``Anomaly detection: A survey,"*ACMComput. Surv.*, vol. 41, no. 3, 2009, Art. no. 15.
[5] W. Eberle and L. Holder, ``Mining for structural anomalies in graph-baseddata," in *Proc. DMin*, 2007, pp. 376_389.
[6] C. C. Noble and D. J. Cook, ``Graph-based anomaly detection," in *Proc.9th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2003,pp. 631_636.
[7] H. Tong and C.-Y. Lin, ``Non-negative residual matrix factorization withapplication to graph anomaly detection," in *Proc. SIAM Int. Conf. DataMining*, 2011, pp. 1_11.
[8] S.Wang, J. Tang, and H. Liu, ``Embedded unsupervised feature selection,"in *Proc. 29th AAAI Conf. Artif. Intell.*, 2015, pp. 470_476.
[9] Z. Lin, M. Chen, and Y. Ma. (2010). ``The Augmented lagrange multipliermethod for exact recovery of corrupted low-rank matrices." [Online].Available: https://arxiv.org/abs/1009.5055.
[10] J. Sun, H. Qu, D. Chakrabarti, and C. Faloutsos, ``Neighborhoodformationand anomaly detection in bipartite graphs," in *Proc. 15th IEEE Int. Conf.Data Mining*, Nov. 2005, p. 8.
[11] A. Patcha and J.-M. Park, ``An overview of anomaly detection techniques:Existing solutions and latest technological trends," *Comput. Netw.*, vol. 51,no. 12, pp. 3448_3470, Aug. 2007.
[12] W. Li, V. Mahadevan, and N. Vasconcelos, ``Anomaly detection andlocalization in crowded scenes," *IEEE Trans. Pattern Anal. Mach. Intell.*,vol. 36, no. 1, p. 18_32, Jan. 2014.
[13] K. Henderson *et al.*, ``It's who you know: Graph mining using recursivestructural features," in *Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2011, pp. 663_671.
[14] F. Keller, E. Müller, and K. Bohm, ``HiCS: High contrast subspaces fordensity-based outlier ranking," in *Proc. ICDE*, Apr. 2012, pp. 1037_1048.
[15] D. Koutra, E. Papalexakis, and C. Faloutsos, ``Tensorsplat: Spotting latentanomalies in time," in *Proc. PCI*, Oct. 2012, pp. 144_149.
[16] J. H. M. Janssens, I. Flesch, and E. O. Postma, ``Outlier detection withone-class classi_ers from ML and KDD," in *Proc. ICMLA*, Dec. 2009,pp. 147_153.
[17] N. A. Heard, D. J. Weston, K. Platanioti, and D. J. Hand, ``Bayesiananomaly detection methods for social networks," *Ann. Appl. Statist.*, vol. 4,no. 2, pp. 645_662, 2010.
[18] J. Tang and H. Liu ``CoSelect: Feature selection with instance selection forsocial media data," in *Proc. SIAM Int. Conf. Data Mining*, 2013, pp. 1_9.
[19] Z. He, X. Xu, and S. Deng, ``Discovering cluster-based local outliers,"*PatternRecognit. Lett.*, vol. 24, nos. 9_10, pp. 1641_1650, 2003.
[20] M. Gupta, J. Gao, C. C. Aggarwal, and J. Han, *Outlier Detection for Tempo-*

*ral Data* (Synthesis Lectures on Data Mining and Knowledge Discovery).San Rafael, CA, USA: Morgan & Claypool, 2014.

[21] J. Tang, Y. Chang, and H. Liu ``Mining social media with social theories:A survey," *ACM SIGKDD Explorations Newslett.*, vol. 15, no. 2,pp. 20_29, 2013.

[22] I. S. Dhillon, S. Mallela, and D. S. Modha, ``Information-theoretic coclustering,"in *Proc. 9th ACM SIGKDD Int. Conf. Knowl. Discovery DataMining*, 2003, pp. 89_98.

# FAKE IMAGE DETECTION WITH ROBUST HASHING

**KUKKALA PAVANI**
PG Scholar, Department of Computer Science,
B.V. RAJU College,
Bhimavaram, W.G.Dt., A.P, India

**K.R.RAJESWARI**
Professor in Computer Science,
B.V. RAJU College,
Bhimavaram,W.G.Dt., A.P, India

**V.BHASKAR MURTHY**
HOD & Associate Professor in Computer Science,
B.V. RAJU College,
Bhimavaram, W.G.Dt., A.P, India

**ABSTRACT**

*With the advancement of technology, the usage of the fake images is the highest, so according to one review, most of the images stored on the server or cloud are transformed or faked. As a result, it is difficult to determine whether the saved images are genuine or not. Therefore, today there are not many systems that can detect whether images are fake or not. Previous histogram-based and feature extraction-based methods have been used to identify fake images. The neural network is the most advanced technology that distinguishes the fake images by examining various features of the image and learning the process of faking an image. As a small step forward in this direction, the proposed system uses features such as dark channel, light channel, RGB channel, and alpha channel. Using the Gaussian distribution, we can analyze the edges of the images to identify fake images. A deep layer analysis is performed using a convolutional neural system along with the fuzzy classification method to improve the probability of detecting false images. Through this research, it is concluded that convolutional neural network (CNN)-based fake colorized image detection performs better than histogram-based and feature extraction-based fake colorized image detection.*

**Keywords:** multiple manipulation techniques, robust hashing, fake-image detection, JPEG compression, GAN.

## 1. INTRODUCTION

Recently, the generative model based on deep learning such as the generative adversarial net (GAN) is widely used to synthesize the photo-realistic partial or whole content of the image and video.Furthermore, recent research of GANs such as progressive growth of GANs (PGGAN)[1] and BigGAN could be used to synthesize a highly photo-realistic image or video so that the human cannot 20 recognize whether the image is fake or not in the limited time. In general, the generative applications can be used to perform the image translation tasks [3]. However, it may lead to a serious problem once the fake or synthesized image is improperly used on social network or platform. For instance, cycleGAN is used to synthesize the fake face image in a pornography video [4]. Furthermore, GANs may be used to create a speech video with the synthesized facial content of any famous politician, causing severe problems on the society, political, and commercial activities. Therefore, an effective fake face image detection technique is desired. In this paper, we have extended our previous study associated with paper ID #1062 to effectively and efficiently address these issues.

In traditional image forgery detection approach, two types of forensics scheme are

widely used: active schemes and passive schemes. With the active schemes, the externally additive signal (i.e., watermark) will be embedded in the source image without visual artifacts. In order to identify whether the image has tampered or not, the watermark extraction process will be performed on the target image to restore the watermark[6]. The extracted watermark image can be used to localize or detect the tampered regions in the target image. However, there is no "source image" for the generated images by GANs such that the active image forgery detector cannot be used to extract the watermark image. The second one-passive image forgery detector–uses the statistical information in the source image that will be highly consistency between different images. With this property, the intrinsic statistical informationcan be used to detect the fake region in the image[7][8]. However, the passive image forgery detector cannot be used to identify the fake image generated by GANs since they are synthesized from the low-dimensional random vector. Nothing change in the generated image by GANs because the fake image is not modified from its original image

Intuitively, we can adopt the deep neural network to detect the fake image generated by GAN. Recently, there are some studies that investigate a deep learning-based approach for fake image detection in a supervised way. In other words, fake image detection can be treated as a binaryclassification problem (i.e., fake or real image). For example, the convolution neural network (CNN) network is used to learn the fake image detector [9]. In [10], the performance of the fake face image detection can be further improved by adopting the most advanced CNN–Xception network [11]. However, there are many GANs proposed year by year. For example, recently proposed GANs such as [1][12][13][14][15][16][3][2] can be used to produce the photo-realistic images. It is hard and very time-consuming to collect all training samples of all GANs. In addition, such a supervised learning strategy will tend to learn the discriminative features for a fake image generated by each GANs. In this situation, the learned detector may not be effective for the fake image generated by another new GAN excluded in the training phase.

In order to meet the massive requirement of the fake image detection for GANs-based generator, we propose novel network architecture with a pairwise learning approach, called common fake feature network (CFFN). Based on our previous approach [5], it is clear that the pairwise learning approach can overcome the shortcomings of the supervised learning-based CNN such as methods in [9][10]. In this paper, we further introduce a novel network architecture combining with pairwise learning to improve the performance of the fake image detection. To verify the effectiveness of the proposed method, we apply the proposed deep fake detector (DeepFD) to identify both fake face and generic image. The primary contributions of the proposed method are two-fold:

• We propose a fake face image detector based on the novel CFFN consisting of several dense blocks to improve the representative power of the fake image.

• The pairwise learning approach is first introduced to improve the generalization property of the proposedDeepFD.

## II. Literature review

Hsu et al. [2] proposed a deep learning-based approach to identify fake images generated by contrast loss using generative adversarial networks (GAN). The author has proposed a fake feature network that is trained using pairwise learning to distinguish the features of real and fake images and outputs the probability of whether the image is real or fake. The proposed system is more powerful than state-of-the-art fake image detectors. Research gap: The proposed system fails when the fake features of a new generator's image are different than in the training phase. At such times, both the counterfeit image and the image detector must be retrained. The next limitation of the proposed technology is a collection of datasets. Yan et al. [3] introduces a system that recognizes a recolored image using the deep discriminative model. The system takes the original image and along with two derived inputs based on the lighting consistency and inter-channel correlation of the original input and outputs the probability of whether or not it is required. The deep discriminatory model is used to improve the performance of the system with greater accuracy. The author proposes an end-to-end deep discriminatory neural network to distinguish natural images from recolored images that captures broader features. Research gap – The system only recognizes recolored images and the design of an effective network architecture is required to look for high-level clues for better discrimination. Read al. [4] developed a system to detect fake colorized images using a neural network. The author analyzes the statistical difference between natural images and corresponding fake images. The author also proposed a simple feature extraction technique to measure the similarity between the normalized histogram distribution of both images. The author conducted experiments by training the dataset created by three state-of-the-art coloring techniques to improve the system's performance and robustness. Guo et al. [5] presented the different types of fakes that should be possible with an image and the different ways that can be used to distinguish the fakes. However, this does not take into account the colorization of grayscale images. This is difficult to detect as there are not numerous systems created to identify miscoloring of images. In this way, to provide an answer to this problem, the creators proposed two different techniques for locating the fake colorized images, the first is the Fake Colorized Image Detection Histogram (FCID-HIST) and the other is the FCID Feature Encoding (FE)-based detection of fake colorized images. These systems are proven to be many times superior to the usual discovery techniques. seen in that way, since it can't tell a fake anyway when the picture is made, but the frame is limited to its understanding. Research Gap – As with any learning technique, a large portion of the identification relies upon the sort of preparing given to the framework and is restricted

Digitization is the process of converting a paper-based handwritten document into electronic format. Here, each document consists of only one character. The electronic conversion is accomplished by using a method whereby a document is scanned and an electronic representation of the original document as an image file format is produced. The author used various scanners for digitization, and the digital image was going for next step that is a preprocessing phase.

Zhao et al. [7] expresses concern about the production of information as vast amounts of information are produced every day. Huge data is a promising field where these gigantic amounts of information can be

processed to derive valuable insights. Whatever the case, the presence of a ton of false, controlled, or produced images has been transcendently corrupted. Consequently, the authors have struggled to isolate the spurious images and distinguish the first image from a particular dataset. The method used by the analysts depends on Convolutional Neural Networks because it is very powerful in the field of image preparation. The system has been tested many times and has provided confirmation that this strategy is one of the most powerful strategies.

Bakhti et al. [8] states that due to the proliferation of web-based life and the scale of image control strategies, there has been an enormous deluge of fake and controlled images. This also leads to spreading many false tricks that generate frenzy and mania. While many calculations have recently been made with the ultimate goal of identifying the imitations, they have failed in a number of situations. The analysts present a system for detecting fake images using the merging of computational and web advances. This process has proven to be more powerful and deeply productive than its partners.

## III. Proposed Methodology

### 1. Feature Extraction

This step extracts basic features from images such as dark channel, light channel, alpha channel and RGB channel. To extract these features, the system looks at an input and iterates through each pixel to get a signed integer (SIGN). Right-shifting the SIGN integer 24 bits and performing an AND operation on a hexadecimal value results in an alpha channel value, in the same way shifting the SIGN integer right 16,8,0 bits and performing an AND operation - Operation with a hexadecimal value results in Red, Green and Blue channel values.

### 2. Edge Estimation

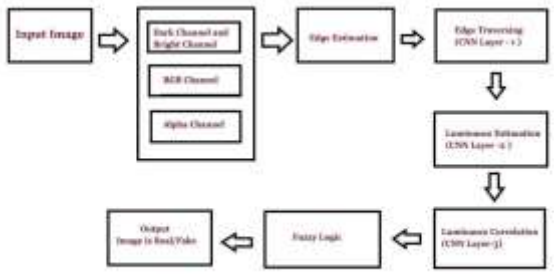
By edges of an image we can tell if the image is fake or real. If the image is real, the edges of the image are intact, while if the image is fake, the values of the edge image will fluctuate. To identify edges, the edge detection algorithm is used together with ROI (Region of Interest) identification and ROI extraction. ROI edges of an image are identified with ROI. The input to the algorithm is a normalized image and the output is a ROI image. PR, PG, PB - Protocol for Red, Green and Blue.

### 3. Edge Traversing

Edge traversal involves traversing edges of the images and noting their luminance value. If the image is real, the luminance value of the edges remains intact. On the other hand, if the image is fake, the luminance value of the edges keeps fluctuating. An edge traversal algorithm is used for this purpose.

### 4. Luminance Correlation

Luminance correlation builds a correlation array by retrieving histogram and luminance values. Another correlation array is created based on the values of the trained image. Both arrays are given as input to Pearson Correlation to check the correlation between both arrays and output the value in the range 0-1.



## IV. MODULES

1. Generate NLBPNet

2. Upload Test Image

3. Classify Picture In Image

1) **Generate NLBPNet** Train & Test Model: in this module we will read all LBP images from LBP folder and then train CNN model with all those images.
2) **Upload Test Image:** In this module we will upload test image from 'testimages' folder. Application will read this image and then extract Deep Textures Features from this image using LBP algorithm.
3) **Classify Picture In Image:** This module apply test image on CNN train model to predict whether test image contains spoof or non-spoof face.

## V. CONCLUSION

In this paper, we have proposed a novel common fake feature network based the pairwise learning, to detect the fake face/general images generated by state-of-the-art GANs successfully. The proposed CFFN can be used to learn the middle- and high-level and discriminative fake feature by aggregating the cross-layer feature representations into the last fully connected layers. The proposed pairwise learning can be used to improve the performance of fake image detection further. With the proposed pairwise learning, the proposed fake image detector should be able to have the ability to identify the fake image generated by a new GAN. Our experimental results demonstrated that the proposed method outperforms other state-of-the-art schemes in terms of precision and recall rate

## 6. REFERENCES

1. Karras, T.; Aila, T.; Laine, S.; Lehtinen, J. Progressive growing of gans for improved quality, stability, and variation. arXiv Preprint, arXiv:1710.10196 2017. 256

2. Brock, A.; Donahue, J.; Simonyan, K. Large scale gan training for high fidelity natural image synthesis. arXiv Preprint, arXiv:1809.11096 2018.

3. Zhu, J.Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired image-to-image translation using cycle-consistent 259 adversarial networks. arXiv Preprint, 2017.

4. AI can now create fake porn, making revenge porn even more complicated,. http://theconversation.com/ai-can-now-create-fake-porn-making-revenge-porn-even-more-complicated-92267, 262 2018.

5. Hsu, C.; Lee, C.; Zhuang, Y. Learning to detect fake face images in the Wild. 2018 International Symposium 264 on Computer, Consumer and Control (IS3C), 2018, pp. 388–391. doi:10.1109/IS3C.2018.00104.

6. H.T. Chang, C.C. Hsu, C.Y.a.D.S. Image authentication with tampering localization based on watermark 266 embedding in wavelet domain. Optical Engineering 2009, 48, 057002.

7. Hsu, C.C.; Hung, T.Y.; Lin, C.W.; Hsu, C.T. Video forgery detection using correlation of noise residue. Proc. of the IEEE Workshop on Multimedia Signal Processing. IEEE, 2008, pp. 170–174.

8. Farid, H. Image forgery detection. IEEE Signal Processing Magazine 2009, 26, 16–25.

9. Huaxiao Mo, B.C.; Luo, W. Fake Faces Identification via Convolutional Neural Network. Proc. of the ACM Workshop on Information Hiding and Multimedia Security. ACM, 2018, pp. 43–47.

10. Marra, F.; Gragnaniello, D.; Cozzolino, D.; Verdoliva, L. Detection of GAN-Generated Fake Images over Social Networks. Proc. of the IEEE Conference on Multimedia Information Processing and Retrieval, 2018, 274 pp. 384–389. doi:10.1109/MIPR.2018.00084.

11. Chollet, F. Xception: Deep learning with depthwise separable convolutions. Proc. of the IEEE conference on 276 Computer Vision and Pattern Recognition 2017, pp. 1610–02357.