IARIA

MMEDIA 2016

The Eighth International Conferences on Advances in Multimedia

ISBN: 978-1-61208-452-7

February 21 - 25, 2016

Lisbon, Portugal

MMEDIA 2016 Editors

Carla Merkle Westphall, Federal University of Santa Catarina, Brazil

Birgit Gersbeck-Schierholz, Leibniz Universität Hannover, Germany

MMEDIA 2016

Forward

The Eighth International Conferences on Advances in Multimedia (MMEDIA 2016), held between February 21-25, 2016 in Lisbon, Portugal, was an international forum for researchers, students, and professionals where to present recent research results on advances in multimedia, and in mobile and ubiquitous multimedia. MMEDIA 2016 brought together experts from both academia and industry for the exchange of ideas and discussions on future challenges in multimedia fundamentals, mobile and ubiquitous multimedia, multimedia ontology, multimedia user-centered perception, multimedia services and applications, and mobile multimedia.

The rapid growth of information on the Web, its ubiquity and pervasiveness, makes the www the biggest repository. While the volume of information may be useful, it creates new challenges for information retrieval, identification, understanding, selection, etc. Investigating new forms of platforms, tools, principles offered by Semantic Web opens another door to enable human programs, or agents, to understand what records are about, and allows integration between domain-dependent and media dependent knowledge. Multimedia information has always been part of the Semantic Web paradigm, but it requires substantial effort to integrate both.

The new technological achievements in terms of speed and the quality expanded and created a variety of multimedia services such as voice, email, short messages, Internet access, m-commerce, mobile video conferencing, streaming video and audio.

Large and specialized databases together with these technological achievements have brought true mobile multimedia experiences to mobile customers. Multimedia implies adoption of new technologies and challenges to operators and infrastructure builders in terms of ensuring fast and reliable services for improving the quality of web information retrieval.

Huge amounts of multimedia data are increasingly available. The knowledge of spatial and/or temporal phenomena becomes critical for many applications, which requires techniques for the processing, analysis, search, mining, and management of multimedia data.

The conference had the following tracks:

- Fundamentals in Multimedia
- Multimedia Applications
- Multimedia services and security

We take here the opportunity to warmly thank all the members of the MMEDIA 2016 technical program committee, as well as all the reviewers. The creation of such a high quality conference program would not have been possible without their involvement. We also kindly thank all the authors that dedicated much of their time and effort to contribute to MMEDIA 2016. We truly believe that, thanks to all these efforts, the final conference program consisted of top quality contributions.

Also, this event could not have been a reality without the support of many individuals, organizations and sponsors. We also gratefully thank the members of the MMEDIA 2016 organizing committee for their help in handling the logistics and for their work that made this professional meeting a success.

We hope MMEDIA 2016 was a successful international forum for the exchange of ideas and results between academia and industry and to promote further progress in the multimedia field. We also hope that Lisbon, Portugal provided a pleasant environment during the conference and everyone saved some time to enjoy the beauty of the city.

MMEDIA 2016 Advisory Committee

Dumitru Dan Burdescu, University of Craiova, Romania
Philip Davies, Bournemouth University, UK
Jean-Claude Moissinac, TELECOM ParisTech, France
David Newell, Bournemouth University, UK
Noël Crespi, Institut Telecom, France
Jonathan Loo, Middlesex University - Hendon, UK
Patrice Rondao Alface, Alcatel-Lucent Bell Labs - Antwerp, Belgium
Trista Chen, Fotolugu Inc, USA
Alexander C. Loui, Kodak Alaris Inc. - Rochester, USA

MMEDIA 2016

Committee

MMEDIA 2016 Advisory Committee

Dumitru Dan Burdescu, University of Craiova, Romania
Philip Davies, Bournemouth University, UK
Jean-Claude Moissinac, TELECOM ParisTech, France
David Newell, Bournemouth University, UK
Noël Crespi, Institut Telecom, France
Jonathan Loo, Middlesex University - Hendon, UK
Patrice Rondao Alface, Alcatel-Lucent Bell Labs - Antwerp, Belgium
Trista Chen, Fotolugu Inc, USA
Alexander C. Loui, Kodak Alaris Inc. - Rochester, USA

MMEDIA 2016 Technical Program Committee

Max Agueh, LACSC - ECE Paris, France
Hakiri Akram, Université Paul Sabatier - Toulouse, France
Musab Al-Hadrusi, Wayne State University, USA
Nancy Alonistioti, N.K. University of Athens, Greece
Giuseppe Amato ISTI-CNR, Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo" - Pisa, Italy
Maria Teresa Andrade, University of Porto / INESC Porto, Portugal
Marios C. Angelides, Brunel University - Uxbridge, UK
Stylianos Asteriadis, University of Maastricht, Netherlands
Ramazan S. Aygun, University of Alabama in Huntsville, USA
Elias Baaklini, University of Valenciennes, France
Andrew D. Bagdanov, Universita Autonoma de Barcelona, Spain
Yannick Benezeth, Université de Bourgogne - Dijon, France
Jenny Benois-Pineau, LaBRI/University of Bordeaux 1, France
Sid-Ahmed Berrani, Orange Labs - France Telecom, France
Steven Boker, University of Virginia - Charlottesville, USA
Fernando Boronat Seguí, Universidad Politecnica de Valencia, Spain
Hervé Bredin, CNRS/LIMSI, France
Marius Brezovan, University of Craiova, Romania
Dumitru Burdescu, University of Craiova, Romania
Nicola Capuano, University of Salerno, Italy
Eduardo Cerqueira, Federal University of Para, Brazil
Damon Chandler, Department of Electrical and Electronic Engineering - Shizuoka University

Copyright Information

Table of Contents

Opinion Mining: A Comparison of Hybrid Approaches

Alex M. G. Almeida
Sylvio Barbon Jr.
Rodrigo A. Igawa

Londrina State University
PR 445 Km 380
Londrina-PR, Brazil
Email: alex.marino@gmail.com
sbarbonjr@uel.br
igawa.rodrigo@gmail.com

Emerson C. Paraiso

Pontifícia Universidade Católica do Paraná
Rua Imaculada Conceição, 1155
Curitiba-PR, Brazil
Email: paraiso@ppgia.pucbr.br

Stela N. Moriguchi

Uberlândia Federal University
Av. João Naves de Avila, 2121
Santa Mônica, Uberlândia-MG, Brazil
Email: stellanm@ufu.br

***Abstract*—Applications based on Opinion Mining and Sentiment Analysis are critical tools for information-gathering to find out what people are thinking. It is one of the most active research areas in Natural Language Processing, and a diversity of strategies and approaches have been published. We evaluate two strategies - Cognitive-based Polarity Identification and Crowd Explicit Sentiment Analysis - and combine them with emoticons and lexicon analysis in a four hybrid models cascade framework. These four approaches were compared to evaluate a suitable method to improve performance over different datasets. We performed experimental tests using the SentiStrength database, which is composed of five public datasets. Results show that emoticons attribution can improve accuracy while combined with Crowd Explicit Sentiment Analysis and Cognitive-based Polarity Identification approaches. In addition, hybrid approaches achieve better precision in case of neutral sentences. Datasets that provide a more informal use of language are suitable for hybrid approaches.**

***Keywords*–Opinion Mining; Sentiment Analysis; Machine Learning.**

I. Introduction

Opinion Mining (OM) and Sentiment Analysis (SA) are fields of studies which analyze opinions, sentiments, and mood, based on textual data. Currently, OM can be related to Information Retrieval (IR) and Machine Learning (ML), which match classification indicators of semantic weight. Although emotions are conceptualized as subjective experiences which are hard to evaluate, for SA they are susceptible to a kind of computing effort accessible from text [1].

Some research fields have arisen due to the increasing use of Web environments, such as forums, blogs, and online social networks (OSNs). By providing ubiquitous information, these environments have encouraged studies concerning people's sentiments, opinions, and mood [2]. Knowledge is obtained from people's written thoughts: in this way, a company can improve its products and services by discovering people's wishes [3].

The OSN, typically, contains a huge volume of opinion in textual format. Additionally, it is a valuable source of information about people's sentiment, but it is not easy for humans to read it due to the large amount of data and the diversity of multimedia characteristics (slangs, emoticons, lack of grammatical accuracy, etc.). Thus, we studied several hybrid combinations such as in [4] associated with ML and lexicon-based approaches. While [4] evaluated only one dataset, our experiments covered five datasets with distinct features: user profile, written profile (varying frequency use of slang, emoticons, writing correction). Furthermore, based on the ML approach, it was possible to achieve accurate results and to identify a high precision combination.

The present paper presents a comparative study of two approaches to OM and SA in the recent literature. Both approaches have been chosen because of their simplicity. The first approach, Crowd Explicit Sentiment Analysis (CESA), is lexicon-based, while Cognitive-based Polarity Identification (CBPI) is a machine learning algorithm. The experiments were performed on five different datasets of web social media content available at the CyberEmotions Consortium [5].

In summary, this paper presents a comprehensive and in-depth critical assessment of both approaches, aiming to highlight the limitations and advantages of each one for predicting opinion polarity in many different datasets. To guide the comparative study, we aim to answer the following questions: a) Which approach is more accurate? b) How do emoticons aid the classification? c) Is there an approach more suitable for a single polarity? d) Does the dataset influence the selection of the approach?

The rest of this paper is organized as follows. Section II gives a brief review of the literature on OM and SA. Then, Section III describes the SentiStrengh Dataset. Next, Section IV describes the experimental settings, followed by Section V, which supplies a detailed descriptions of the approaches used in this paper. The following Section VI describes the results, and Section VII presents some conclusions and suggestions for future work.

II. Related Work

Prior to classifying opinions, extracting opinions is also a concern of OM and SA. In [6], the authors report experiments addressing feature-based opinion extraction and achieved good results limited to 40 pre-defined examples. In [7], classic Text Mining pre-processing techniques based on OM were studied and it was concluded that this kind of technique should be adapted in order to correctly clean the noisy text from platforms such as Twitter.

ISBN: 978-1-61208-452-7

A more recent paper about Opinion Retrieval, a subarea of OM, used YouTube, where the goal was to obtain the target of an opinion as being either the video or the product [8].

SA is an important research area for audio and video. Recently [9] showed the importance of several feature extraction methods in experiments with EmoDB Dataset [10], which Suport Vector Machines (SVM) classification improved 11% in accuracy. In [11], J.G. Ellis et al. focus on predicting the evoked emotion at particular times within a movie trailer. The main idea is to learn a mid-level feature representation for multimedia content that is grounded in machine detectable concepts and then model human emotions based on this representation.

Regarding SA, the literature can be divided into two main approaches: Lexicon and Machine Learning. The first one mainly relies on a sentiment lexicon, i.e., a set of known and predefined terms or phrases that represent emotions, e.g., Opinion Finder available on [12].

On the other hand, Machine Learning approaches rely on an initial set of pre-labeled documents, opinions, or terms, to automatically extract features for further classification [13].

Within Lexicon-based SA, there are two more subdivisions of approaches, according to [13], [14]: Dictionary and Corpus-based. The second usually concerns a more dynamic set of words rather than fixed dictionaries to represent emotions. For example, in [15], the goal was to retrieve a new and adapted lexicon from a specific domain.

Also, in [16], it was reported that only one lexicon in a reference language should be necessary to perform a multi-language SA. In [17], the authors successfully analyzed the behavior of sports fans during FIFA 2014 World Cup on Twitter.

Dictionary-based approaches have also used a set of terms to be updated according to context as shown in [18], [19], the analysis of stock prices, or of the emergence of political topics, all based on the news. One last example of a dictionary-based solution is CESA, exposited in [20], where a dictionary was shown to be useful in combination with other tools.

Unlike Lexicon-based approaches, Machine Learning applied to OM and SA mainly depends on its labeled corpus to extract features in order to classify opinions. Examples are shown in [21], [22], in which SVM was applied to monitor not only the products, but also features that would describe or classify opinions in multiple domains.

Another case of the use of Machine Learning is to classify opinion polarity, based on sentence features [23]. Probabilistic classifiers (e.g., Naive Bayes) in [2], [24] also performed well at inferring the polarity of tweets, in a simple way. In this paper, experiments as with [24], have been used as an ML approach for comparison of lexicon-based approaches.

III. The SentiStrength Dataset

The SentiStrength [25] database consists of six datasets: Digg, BBC forum, Runners World forum, YouTube, Twitter, and MySpace. The dataset in its original form includes three fields:

1) positive strength
2) negative strength
3) message

TABLE I. Sentiment Strength x Sentiment Label

PS	NS	Message	SL
3	-1	I am so happy, #SherlockHolmes incredible soundtrack! Glad to see others appreciated. Now if it wins, crossing my fingers.	positive
1	-1	Grow on Twitter with Tweet Automator http://bit.ly/dwxFHu	neutral
1	-2	#4WordsOnObamasHand Don't Say The N-Word	negative

To make it possible to use datasets for sentiment classification, we reassigned all messages in datasets with sentiment labels (positive, negative, neutral) rather than sentiment strengths (positive and negative strengths). In Table I, the negative strength (NS) is a number between -1 and -5, and the positive strength (PS) is a number between 1 and 5.

To obtain sentiment label (SL) we followed two rules: neutral if the difference between the negative absolute value and the positive value is 0, and positive if the ratio of positive values to the negative values is bigger than 1.5; if not, it is negative [26].

IV. Experimental Settings

To perform this experiment, we used a reassigned dataset, as described in Section III, called the sample. The sample was composed of labeled messages (positives, neutrals, and negatives). Table II presents some sample characteristics: number of negative and positive emoticons, number positive and negative sentiment terms, and the number of messages for each dataset. Figure 1 shows the algorithm of the experimental procedure. We partitioned it into four tasks: Acquisition, Pre-processing, Training and Classification.

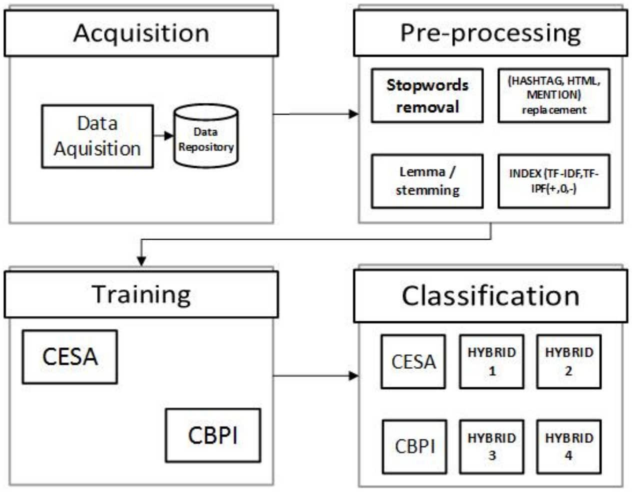


Figure 1. Complete Experimental Process Diagram

For the Acquisition task, we used SentiStrength Database where messages serve as input to the pre-processing task. The MySpace Dataset was ignored because is a deprecated OSN.

The pre-processing task performs intensive processing steps for each message and then sends it to the subsequent task. The pre-processing task consists of the following steps:

1) Stop words removal
2) Tags replacement ("@", "#")

TABLE II. DATASETS SPECIFICATION

Dataset	Emoticons		SWN Terms		Messages
	NEG	POS	NEG	POS	
YOUTUBE	46	277	595	2012	3407
TWITTER	130	427	556	1450	4127
RW1046	41	181	678	1758	1042
DIGG	10	28	471	742	1042
BBC	7	13	753	1062	1042

3) Stemming
4) Indexing by Term Frequency-Inverse Document Frequency (TF-IDF) and Term Frequency-Inverse Polarity Frequency (TF-IPF)

For the Training task, we focused on building a classifier based on the input vector plus the answer vector to generalize the knowledge and finally predict the messages. In the main CESA and CBPI, we used a 10-fold cross-validation. With regard to the Classification task, accuracy refers to the ability of the classifier to predict the class label correctly for a new set of data.

V. THE COMPARED APPROACHES

This article includes different approaches of previous research. In these studies, the main issues are classification accuracy, data sparsity with insufficient data or very few useful labels in the training set, and a high percentage of sentences incorrectly classified as neutral [4]. For this research, we chose CESA due to the fact that it is simple to implement [20] and highly accurate for negative and positive strengths.

The CBPI approach was chosen because it is a simple solution and has one of the best neutral classification results. The challenge of hybrid approaches is to combine these techniques with emoticon interpretation in order to improve the accuracy of the results. Moreover, a hybrid approach can reduce the dependency of ML [27] on SentiWordNet (SWN) [28] because a prior polarity attribution diminishes the influence of the training set on an ML solution, which is good for low-quality datasets.

A. *Emoticon Attribution*

Emoticon attribution is used in a simple way. Attribution is done based on a list of emoticons. It uses regular expressions, based on Table III, to detect the presence of emoticons which are then classified as positive or negative and reduces the dependency on machine learning [27].

TABLE III. EMOTICONS POLARITY

Emoticon	Polarity
:-) :) :o) :] :3 :c) :N =] 8) =) :} :^) :)	positive
:-D :D 8-D 8D x-D xD X-D XD =-D =D=-3 =3 B^D	positive
>:[:-(:(:-c :c :-<:<:-[:[:{	negative
>:\>:/ :-/ :-. :/ :\=/ =\:L =L :S >.<	negative

B. *SentiWordNet*

SWN uses a list of positive and negative words, as shown in Table IV, to check the sentiments for each term in the message. After pre-processing, for each word, the sentiment score is found and the final polarity is given by the sum of each sentiment score.

TABLE IV. POSITIVE AND NEGATIVE POLARITY OF WORDS

Sentiment Words	Polarity
better good well great happy free best lucky safe fine ready big strong special normal	positive
bad guilty sick sad lost tired ill stupid weird horrible wrong terrible hurt worse empty uncomfortable	negative

Let a document *D* be defined as a set of messages, as follows:
$D = \{m_1, m_2, m_3, ..., m_i\}$.

Let *m* defined as a set of words for each message, as follows:
$m = \{w_1, w_2, w_3, ..., w_j\}$.

Let *PW* be defined as a set of positive words and *NW* defined as negative words as
PW=$\{$set of positive words$\}$ and
NW=$\{$set of negative words$\}$.

For each word in a message the score is calculated by (1).

$$S(w_j) = \begin{cases} 1 & (w_j \in m_i) \land (m_i \in D) \land (w_j \in PW); \\ -1 & (w_j \in m_i) \land (m_i \in D) \land (w_j \in NW); \\ 0 & (w_j \in m_i) \land (m_i \in D) \land (w_j \notin (NW \cup PW)). \end{cases} \tag{1}$$

The final message polarity of a message obtained by SWN is given (2).

$$P(m_i) = \begin{cases} positive & \sum_{k=1}^{j} S(w_k) > 0; \\ negative & \sum_{k=1}^{j} S(w_k) < 0. \end{cases} \tag{2}$$

C. *Crowd Explicit Sentiment Analysis*

The lexicon-based OM used in this research is CESA [20]. Based on Explicit Semantic Analysis (ESA), the main idea consists in using a set of documents as a matrix to represent the meaning of the concepts obtained from a *corpus* [29]. The main improvement of CESA over ESA is grounded in forming a vector of sentiments.

As with any other lexicon-based approach, CESA requires a *Lexicon v*; in this paper, we use WeFeelFine [30]. The first step of CESA consists in acquiring the dataset, followed by pre-processing. These pre-processing techniques result in a sentiment vector—to be manually labeled—along with the initial *corpus* and the TF-IDF.

The matrix M_{mn}, the key part of the CESA, results from a combination of the sentiment vector and the *corpus*. m represents the size of the *corpus* and n the number of sentiments obtained by the pre-processing phase. We perform a modified version of the CESA: in our version, the part of speech tagger was not employed, due to the results shown in [7], which concluded that the classical part of speech tagger should be adapted when applied to an informal *corpus* like an OSN.

D. *Cognitive-based Polarity Identification*

The Machine Learning approach used in our experiments is found in [24]. As specified in its own presentation, the Cognitive-based Polarity Identification prioritizes simplicity and uses a Bayesian Classifier along with a TF-IPF for the feature extraction. By this time, there should be three sets: a set of positive posts, another of negative posts, and a set of neutral posts.

From there, the procedure takes into consideration an adapted TF-IDF which the authors call TF-IPF. The TF measures how frequently a term *t* occurs in the *i*th polarity document, while the Inverse Polarity Frequency measures how important *t* is for that polarity class. Then, the polarity score $S_i(t)$ (where $i = 1, 2, 3$) considering positive, negative, and neutral is associated to the term t for the polarity i is $S_{class_i}(t) = TF_i(t) * IPF(t)$.

As a consequence, the terms containing higher values of $S_{class_i}(t)$ are considered features. Following the same threshold used in [24], we considered as document feature any term with a $S_{class_i}(t)$ greater than 0.5.

By applying Naive Bayes, the aim is to predict the class c_i from the probability that the document D, represented as a set of m features f_k where $(k = 1, 2, 3, .., m)$ belongs to that class $p(c_i|D)$, given by (3) where $p(f_k|c_i)$ is the probability that the features belong jointly to the class c_i (A is a normalization factor). Lastly, the classifier predicts the class i of the document D presenting the highest probability.

$$p(c_i|f_1, f_2, .., f_m) = \frac{1}{A} p(c_i) \prod_{k=1}^{m} p(f_k|c_i) \qquad (3)$$

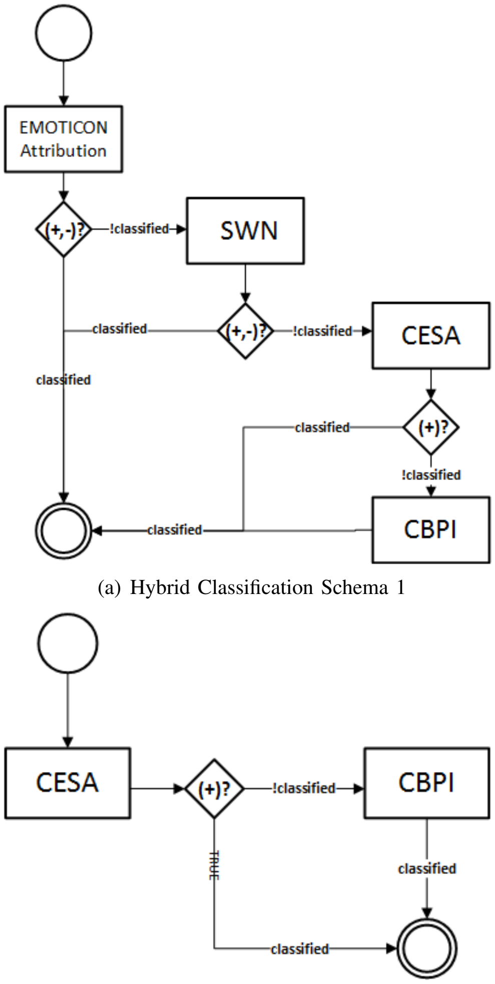
E. Hybrid Schema 1

Hybrid Classification Schema 1, as shown in Figure 2(a), is composed of four steps:

1) Emoticon Attribution
2) SWN
3) CESA
4) CBPI

Emoticon Attribution and SWN are baseline filters, and the CESA and CBPI are machine learning methods. To define Emoticon Attribution as the first step on baseline filters, we considered the fact that the presence of an emoticon in a microblog message represents a sentiment that extends to the whole message [2]. For ML methods, we defined CESA as the first step because it is good for negative and positive classification while CBPI is one of the best methods for neutral rating. To perform the classification, each message is subjected to the four indicated steps in order. The step Emoticon Attribution classifies a particular message as positive or negative verifying the existence or not of emoticons. When subjected message does not contain any emoticon, then the message is processed by SWN which classifies it only as positive or negative, as shown in V-B. Once all filtering steps are performed and the message is not classified yet (as positive or negative), then it is processed by CESA.

When a subjected message comes to CESA step it emphasizes positive messages, in other words, it is being verified if the message is positive or not. If the message is not classified as positive, then it is finally treated by CBPI, which will classify it as positive, neutral, or negative. Hybrid 1 is done using the three scores as given in (4).



(a) Hybrid Classification Schema 1

(b) Hybrid Classification Schema 3

Figure 2. Hybrid classifications schemas

$$P(m) = \begin{cases} positive & (S_E > 0) \lor (S_E = 0 \land S_S > 0) \lor \\ & (S_E = 0 \land S_S = 0 \land S_D > 0) \lor \\ & (S_E = 0 \land S_S = 0 \land S_D \leq 0 \land S_M > 0); \\ negative & (S_E < 0) \lor (S_E = 0 \land S_S < 0) \lor \\ & (S_E = 0 \land S_S = 0 \land S_D \leq 0 \land S_M < 0); \\ neutral & (S_E = 0 \land S_S = 0 \land S_D = 0 \land S_M = 0). \end{cases} \qquad (4)$$

where S_E, S_S, S_D and S_M are Emoticon Attribution, SWN, CESA and CBPI respectively.

F. Hybrid Schema 2

Hybrid 2 performs the same baseline filters as Hybrid 1. Hybrid 2 differs from Hybrid 1 in the CESA step, which emphasizes negative messages. The next step, CBPI, performs the same way as Hybrid 1 and the message is classified as positive, neutral or negative. Hybrid 2 is done using the three scores as given in (5).

$$P(m) = \begin{cases} positive & (S_E > 0) \lor (S_E = 0 \land S_S > 0) \lor \\ & (S_E = 0 \land S_S = 0 \land S_D \geq 0 \land S_M > 0); \\ negative & (S_E < 0) \lor (S_E = 0 \land S_S < 0) \lor \\ & (S_E = 0 \land S_S = 0 \land S_D < 0) \lor \\ & (S_E = 0 \land S_S = 0 \land S_D \geq 0 \land S_M < 0); \\ neutral & (S_E = 0 \land S_S = 0 \land S_D = 0 \land S_M = 0). \end{cases} \qquad (5)$$

G. Hybrid Schema 3

Hybrid classification schema 3 does not perform baseline filters, as shown in Figure 2(b).

All pre-processed messages are directly evaluated by CESA emphasizing positive message and, in case the current message is not positive, it is handed on to CBPI, which will classify the message as positive, neutral, or negative. Hybrid 3 is done using the three scores as given in (6).

$$P(m) = \begin{cases} positive & (S_D > 0) \vee (S_D \leq 0 \wedge S_M > 0); \\ negative & (S_D \leq 0 \wedge S_M < 0); \\ neutral & (\mathrm{S}_D \leq 0 \wedge S_M = 0). \end{cases} \quad (6)$$

H. Hybrid Schema 4

In a similar way to Hybrid 3, Hybrid 4 differs in emphasizing negative messages. All messages not classified at CESA as negative are treated by CBPI so as to be finally classified as positive, neutral or negative. Hybrid 4 is done using the three scores as given in (7).

$$P(m) = \begin{cases} positive & (S_D \geq 0 \wedge S_M > 0); \\ negative & (S_D < 0) \vee (S_D \geq 0 \wedge S_M < 0); \\ neutral & (\mathrm{S}_D \geq 0 \wedge S_M = 0). \end{cases} \quad (7)$$

VI. Analysis Results and Discussion

To evaluate the proposed hybrid approaches, we used confusion matrix, precision, accuracy and a ranking order scale. The division of true positive against both true positive and false positive defines precision, which is denoted as:

$$Precision_A = \frac{TP_A}{TP_A + FP_{BA} + FP_{CA}}$$

where TP_A represents the number of right predictions for class A while FP_{BA} are class B incorrectly classified as A, and FP_{CA} are class C incorrectly classified as A.

The accuracy is the proportion of true results (both true positives and true negatives) among the total number of cases examined which is denoted as:

$$Accuracy = \frac{Tpos + Tneg + Tneu}{Tpos + Fpos + Tneg + Fneg + Tneu + Fneu}$$

where Tpos, Tneg and Tneu are respectively true positives, true negatives and true neutrals while Fpos, Fneg and Fneu are false positive, false negative and false neutral respectively.

The rank is a comparative scale technique where 1 is assigned to the best accuracy obtained from an approach per dataset, and 6 is assigned to the worst accuracy. This rank is an ordinal scale that describes accuracies performance approaches, but does not revel distance between approaches. We use a final ranking that is calculated by performance average of each method over all datasets.

Our discussion of the results starts with the first question presented in Section I: "a) Which approach is more accurate?" Figure 3 presents a box plot of each approach along with their respective accuracies. The box plots show that the approaches Hybrid 1 and Hybrid 2 presented the highest average accuracy (75.00% and 77.00%) along with the highest values of quartile 1 (Q1) (65.50% and 65.75% accuracy).

In another view, the CESA approach presented the highest accuracy but in terms of stability it was not the best method because CESA's box has a similar size to Hybrid 4 box that is clearly the most unstable. Although CESA presented a median value close to Hybrid 1 and Hybrid 2, it is very important to highlight that not only medians and maximum values of accuracies should be taken into consideration. If one of these methods should be selected, Hybrid 1 and Hybrid 2 are the most appropriate because of their stability presented in the box plot.

Thus, when maintaining stability is a matter of greater importance than obtaining the best accuracy in real scenario, we recommend Hybrid 1 or Hybrid 2. This is our conclusion regarding the first question once CESA presented either high and low results in terms of accuracy.

TABLE V. Accuracies and Ranking on the Datasets

Dataset	CESA	CBPI	Hybrid 1	Hybrid 2	Hybrid 3	Hybrid 4
BBC	83.13% (1)	50.39% (5)	76.74% (3)	77.34% (2)	48.64% (6)	67.67% (4)
Digg	84.15% (1)	43.91% (5)	74.63% (3)	78.47% (2)	41.59% (6)	63.42% (4)
Runners	58.95% (3)	40.76% (5)	68.24% (2)	69.71% (1)	46.47% (4)	32.65% (6)
Twitter	53.19% (3)	41.94% (4)	58.26% (1)	53.33% (2)	35.07% (5)	27.25% (6)
Youtube	70.66% (3)	42.12% (5)	76.99% (2)	77.78% (1)	58.11% (4)	41.30% (6)
Ranking	2.2	4.8	2.2	1.6	5.0	5.2

To answer the second question, "b) How do emoticons aid the classification?", we performed experiments concerning the Hybrid approaches, and the results are shown in Figure 4(a). Note that Hybrid 1 and Hybrid 2 both take into consideration emoticons while Hybrid 3 and Hybrid 4 do not take advantage of that aspect. It is possible to see that, apart from which dataset the approach was performed on, considering emoticons always yielded higher results, as we can see by the fact that the bar plots show better results than the line plots for all datasets.

Figure 4(b) illustrates our results regarding the question: "c) Is there an approach more suitable for a specific polarity?" This question is especially addressed in a case of discovering a specific polarity that is important and, therefore, making an analysis of the precision concerning one specific class would also be necessary.

In Figure 4(b), it is notable that the prediction of neutral opinions is not a problem for either Hybrid 1 and Hybrid 2. However, if it is necessary to achieve higher results about positive or negative, then Hybrid 1 must be selected to predict neutral and negative opinions (78.62%) while Hybrid 2 must be chosen to predict neutral or positive opinions (82.33%).

The last question in Section I: "d) Does the dataset influence the selection of an approach?" is illustrated by Table V. Each row in the table corresponds to accuracy, in percentage, and ranking of accuracies per dataset. The last row in the table represents the average ranking per approach. The average ranking shows that Hybrid 2 (1.6) is the best average method for evaluated datasets followed by Hybrid 1 and CESA (2.2). It is important to note that in datasets, such as BBC and Digg, that provide a more structured text than the Twitter scenario, the CESA approach performs with satisfactory results (e.g., 83.13% and 84.15%, respectively and ranking 1).

On the other hand, Hybrid 1 and 2 approaches achieve better results in datasets like Twitter (ranking 1 and 2 respectively) and YouTube (ranking 2 and 1 respectively) by taking emoticons into consideration. For example, the row for YouTube Dataset of Table V shows the accuracies on Hybrid

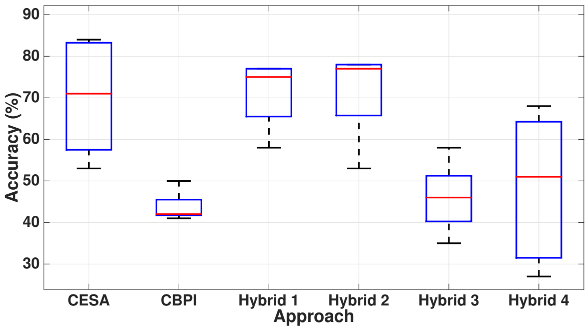
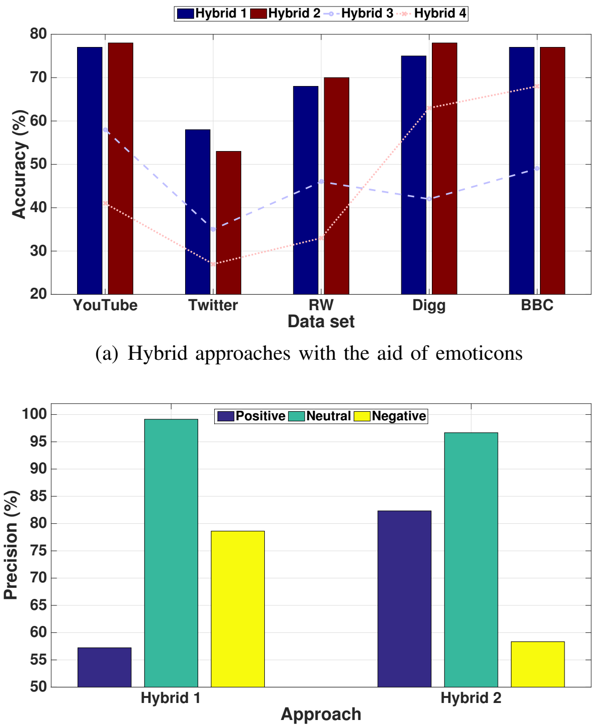


Figure 3. Box plot accuracies of each approach over all datasets

1 and Hybrid 2 achieved 76.99% and 77.78% (ranking 2 and 1 respectively) while the CESA approach achieved 70.66% (ranking 3).



(a) Hybrid approaches with the aid of emoticons

(b) Polarity Tendency on Hybrid 1 and Hybrid 2

Figure 4. Overview of the hybrid approaches Accuracy and Precision

Hybrid 3 and Hybrid 4 presented lower accuracy results than Hybrid 1 and Hybrid 2. As discussed in Figure 4(a), this results from the fact that they do not take advantage of emoticon analysis. Therefore, datasets providing a more informal use of language should indicate the selection of an approach concerning informal features like emoticons is a better choice.

The CBPI approach achieved lower results in terms of accuracy in the datasets used in our experiments. This can be justified by the fact that a better threshold from TF-IPF was not found and the words used as features on Naive Bayes might not have described the data properly. Still, CBPI is the most stable method in that it achieves similar results in different sets, as shown in Figure 3.

VII. Conclusion and Future Work

We provided in this paper a comparative study using reported datasets with distinct characteristics. Our results have shown that the content of datasets with an informal use of language is better classified by hybrid schemes reinforced with emoticon attribution and SWN.

Still, as to the emoticon attribution and SWN aid for the approaches Hybrid 1 and Hybrid 2, we noted that the precision is improved, reaching more than 95%, on single neutral classification independently of the dataset, whereas positive and negative single classification alternated good results (in precision terms) respectively.

Hybrid approaches 1 and 2 resulted in more stable results in terms of accuracy among different datasets, which constitutes the relevant contribution of this paper. For future research, we suggest focusing on the computational cost and specifically on a deeper research into sarcasm, which we believe to be intrinsically connected to OM as a great challenge.

References

[1] B. Liu and L. Zhang, “A survey of opinion mining and sentiment analysis,” in Mining Text Data. Springer, 2012, pp. 415–463.

[2] A. Pak and P. Paroubek, “Twitter as a corpus for sentiment analysis and opinion mining.” in LREC, 2010.

[3] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau, “Sentiment analysis of twitter data,” in Proceedings of the Workshop on Languages in Social Media. Association for Computational Linguistics, 2011, pp. 30–38.

[4] F. H. Khan, S. Bashir, and U. Qamar, “Tom: Twitter opinion mining framework using hybrid classification scheme,” Decision Support Systems, vol. 57, 2014, pp. 245–257.

[5] “Cyberemotions consortium,” [accessed October-2015]. [Online]. Available: http://sentistrength.wlv.ac.uk/

[6] F. L. Cruz, J. A. Troyano, F. Enríquez, F. J. Ortega, and C. G. Vallejo, “Long autonomy or long delay? the importance of domain in opinion mining,” Expert Systems with Applications, vol. 40, no. 8, 2013, pp. 3174–3184.

[7] G. Petz, M. Karpowicz, H. Fürschuß, A. Auinger, V. Stříteskỳ, and A. Holzinger, “Computational approaches for mining user’s opinions on the web 2.0,” Information Processing & Management, vol. 50, no. 6, 2014, pp. 899–908.

[8] A. Severyn, A. Moschitti, O. Uryupina, B. Plank, and K. Filippova, “Multi-lingual opinion mining on youtube,” Information Processing & Management, 2015.

[9] E. S. Erdem and M. Sert, “Efficient recognition of human emotional states from audio signals,” in Multimedia (ISM), 2014 IEEE International Symposium on. IEEE, 2014, pp. 139–142.

[10] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, “A database of german emotional speech.” in Interspeech, vol. 5, 2005, pp. 1517–1520.

[11] J. G. Ellis, W. S. Lin, C.-Y. Lin, and S.-F. Chang, “Predicting evoked emotions in video,” in Multimedia (ISM), 2014 IEEE International Symposium on. IEEE, 2014, pp. 287–294.

[12] “Opinion finder,” [accessed October-2015]. [Online]. Available: http://mpqa.cs.pitt.edu/opinionfinder/

[13] J. Serrano-Guerrero, J. A. Olivas, F. P. Romero, and E. Herrera-Viedma, “Sentiment analysis: A review and comparative analysis of web services,” Information Sciences, vol. 311, 2015, pp. 18–38.

[14] C. Potts, “Sentiment symposium tutorial,” in Sentiment Symposium Tutorial. Acknowledgments, 2011.

[15] S. Park, W. Lee, and I.-C. Moon, “Efficient extraction of domain specific sentiment lexicon with active learning,” Pattern Recognition Letters, vol. 56, 2015, pp. 38–44.

[16] A. Hogenboom, B. Heerschop, F. Frasincar, U. Kaymak, and F. de Jong, “Multi-lingual support for lexicon-based sentiment analysis guided by semantics,” Decision support systems, vol. 62, 2014, pp. 43–53.

[17] Y. Yu and X. Wang, “World cup 2014 in the twitter world: A big data analysis of sentiments in us sports fans tweets,” Computers in Human Behavior, vol. 48, 2015, pp. 392–400.

[18] X. Li, H. Xie, L. Chen, J. Wang, and X. Deng, “News impact on stock price return via sentiment analysis,” Knowledge-Based Systems, vol. 69, 2014, pp. 14–23.

[19] S. Rill, D. Reinel, J. Scheidt, and R. V. Zicari, “Politwi: Early detection of emerging political topics on twitter and the impact on concept-level sentiment analysis,” Knowledge-Based Systems, vol. 69, 2014, pp. 24–33.

[20] A. Montejo-Ráez, M. Díaz-Galiano, and L. Ureña-López, “Crowd explicit sentiment analysis,” Knowledge-Based Systems, 2014.

[21] M. Zimmermann, E. Ntoutsi, and M. Spiliopoulou, “Discovering and monitoring product features and the opinions on them with opinstream,” Neurocomputing, vol. 150, 2015, pp. 318–330.

[22] M. R. Saleh, M. T. Martín-Valdivia, A. Montejo-Ráez, and L. Ureña-López, “Experiments with svm to classify opinions in different domains,” Expert Systems with Applications, vol. 38, no. 12, 2011, pp. 14 799–14 804.

[23] J. M. Chenlo and D. E. Losada, “An empirical study of sentence features for subjectivity and polarity classification,” Information Sciences, vol. 280, 2014, pp. 275–288.

[24] E. D’Avanzo and G. Pilato, “Mining social network users opinions’ to aid buyers shopping decisions,” Computers in Human Behavior, 2014.

[25] M. Thelwall, K. Buckley, and G. Paltoglou, “Sentiment strength detection for the social web,” Journal of the American Society for Information Science and Technology, vol. 63, no. 1, 2012, pp. 163–173.

[26] H. Saif, M. Fernández, Y. He, and H. Alani, “Evaluation datasets for twitter sentiment analysis: a survey and a new dataset, the sts-gold,” 2013.

[27] J. Read, “Using emoticons to reduce dependency in machine learning techniques for sentiment classification,” in Proceedings of the ACL Student Research Workshop. Association for Computational Linguistics, 2005, pp. 43–48.

[28] S. Baccianella, A. Esuli, and F. Sebastiani, “Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining.” in LREC, vol. 10, 2010, pp. 2200–2204.

[29] E. Gabrilovich and S. Markovitch, “Computing semantic relatedness using wikipedia-based explicit semantic analysis.” in IJCAI, vol. 7, 2007, pp. 1606–1611.

[30] S. D. Kamvar and J. Harris, “We feel fine and searching the emotional web,” in Proceedings of the fourth ACM international conference on Web search and data mining. ACM, 2011, pp. 117–126.

Automatic Labelling of Seeds Based on Saliency Detection and Edge Detection for Image Segmentation

Cheng-Mao Wu
Department of Computer Science
National Tsing Hua University
Hsinchu, Taiwan
email:s101062613@m101.nthu.edu.tw

Long-Wen Chang
Department of Computer Science
National Tsing Hua University
Hsinchu, Taiwan
email:lchang@cs.nthu.edu.tw

***Abstract*— In computer vision, image segmentation transforms an input image into a more meaningful form which is easier to analyze. It can be used in the applications such as medical imaging, object detection, face recognition, etc. Generally, image segmentation can be distinguished as supervised and unsupervised categories. The result of supervised image segmentation is greatly affected by a user. Therefore, we propose an unsupervised method of image segmentation. We use saliency detection to label some informative and significant parts of the image, and then, we apply edge detection to label some details of the image and use the labelled image for image segmentation by Kim's method. In this way, we can automatically label the seeds to get the scribble and then segment the image into the foreground and the background. The simulation results show that our method is feasible for image segmentation.**

***Keywords*- segmentation;supervised segmentation; unsupervised segmentation; saliency detecion; edge detection.**

I. INTRODUCTION

In computer vision, image segmentation is the process of partitioning an image into several segments. The goal of image segmentation is to transform the input image into a more meaningful form which is easier to analyze. How to detect the objects which humans can recognize in the image is a big issue in image segmentation. Generally, image segmentation can be distinguished as supervised and unsupervised methods.

Supervised image segmentation [1]-[3] needs the user to label some seeds as scribbles. However, the result is greatly affected by the user. The result of segmentation image is shown in Figure 1 while using different scribbles in the Kim's algorithm [3]. Besides, supervised image segmentation is hard to evaluate the interactive time. Therefore, we propose an unsupervised method of image segmentation, which can't be affected by users.

Because the supervised image segmentation methods have some above-mentioned drawbacks. Thus, our motivation is to improve the supervised method's disadvantage and make our result of the unsupervised method close to that of the supervised method. Our goal is to transform the supervised method to the unsupervised method. We use saliency detection [7] and edge detection [8] to automatically label the seeds to get the scribble, and then we apply Kim's method to segment the image into the foreground and the background.

Recently, there are a lot of saliency detection method [4]-[7]. Moreover, we also apply Canny edge detection [8] to detect the edge because the method is simple and efficient. By saliency detection [7] and edge detection [8], we can automatically label seeds for image segmentation. Finally, we apply our scribble image with Kim's method [3] to get the final segmentation result. We will compare our method with Kim's method (supervised method) and the method of Donoser et. al [10] which is also unsupervised and based on saliency. The results show that our approach is good for image segmentation. In the section 2, we will briefly discuss some related works. In section 3, we will address our proposed algorithm. In section 4, we will show some experiment results and finally, we give our conclusion.

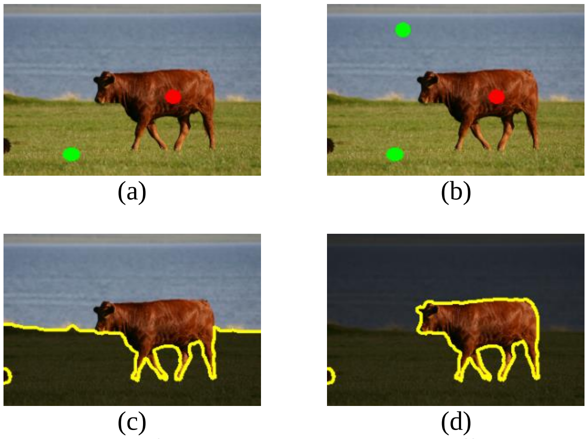


Figure 1. The result of segmentation image while using different scribbles in the Kim's algorithm. (a) (b) The different scribbles labeled by different users. (c) The result of (a). (d) The result of (b)

II. RELATED WORKS

Image segmentation can be categorized into many categories: thresholding methods, clustering methods, region-based methods, edge-based methods, etc. The thresholding methods is the simplest method of image segmentation. The input image is transformed into a gray-scale image, and then segmented by a threshold value to get a binary image [11]. The main concept of clustering [12],[13] is to determine which components of a data set naturally "belong together". The region-based methods are widely used as well [14], [15]. Because segmentation consists on partitioning an image into a set of connected regions, we can find homogeneous regions according to a specific criterion (intensity value, texture). In addition, segmentation can also be done by edge detection techniques [8]. Donoser et al. [10] proposed another method which is unsupervised and based on saliency. The method automatically find salient regions first, and then focus on one different salient part of the image each time; the method finally, merge all obtained results into a composite segmentation.

We will compare the result of our method with that of Donoser et al. [10] because it is unsupervised and based on saliency as well. Furthermore, we will also compare the result of our method with the result of Kim's method (supervised method) [3]. According to the results, it shows that our result is good and very close to the result of Kim's method.

III. Proposed Method

The main idea of our method is labeling the seeds automatically and efficiently. To reduce the user's influence, we use our method to get the scribble of the image. The scribble provides the seeds to the method of nonparametric higher-order learning for segmentation [3].

First, given an input image, we detect the saliency of the image by Yang's method [7]. After saliency detection, we can automatically label the seeds. It contains three steps:

1. The foreground seeds and the background seeds by the foreground and the background threshold finding.
2. Locate the background seeds with four side examination.
3. Combine the seeds from step 1 and step 2 with the seeds generated by edge and saliency detection.

By these three steps, we can automatically and efficiently label the seeds to get a scribble image.

In the step of FG/BG threshold finding, we calculate the number of each saliency value and get the histogram of pixel's saliency value. We find that the background region is at least occupy 1/b area of an image, where b is a parameter (i.e., the lowest n_l saliency pixels of an image). Thus, we can get a proper background threshold (B_T) as

$$\sum_{sv=0}^{B_T} \mathrm{Num}(sv) \geq \mathrm{n}_l = \frac{N}{b}, \tag{1}$$

where N is total number of pixels in the image, sv is saliency value (sv = 0,1,2, ..., 255), Num(sv)：the number of pixels whose saliency value is sv.

Similarly, we find that the foreground region is at least occupy 1/f area of an image, where f is a parameter (i.e., the highest n_h saliency pixels of an image). Thus, we can get a proper foreground threshold (F_T) as

$$\sum_{sv=0}^{F_T} \mathrm{Num}(255 - sv) \geq \mathrm{n}_h = \frac{N}{f}, \tag{2}$$

where N is total number of pixels in the image, sv is saliency value (sv = 0,1,2, ⋯, 255), Num(sv)：the number of pixels whose saliency value is sv.

Then, we label the FG/BG and remain some undetermined parts. For each pixel x_i, and its saliency value $\mathrm{S}(x_\mathrm{i})$, if $\mathrm{S}(x_\mathrm{i})$ is larger than F_T or is equal to F_T, label it as the foreground pixels (red). If $\mathrm{S}(x_\mathrm{i})$ is smaller than B_T or is equal to B_T, label it as the background pixels (green). If $\mathrm{S}(x_\mathrm{i})$ is larger than $B_T$but smaller than F_T, do nothing (undetermined). Thus, we can get 3 parts, the foreground, the background and the undetermined.

In addition to find proper the foreground/background thresholds, we also examine four sides of the image. By step 1, we've already known where the foreground is approximately located. We want to check the left, right, top, bottom regions, shown in Figure 2 and then label some background seeds.

We locate 4 regions $r_{s(l)}$, $r_{s(r),} r_{s(t)}, r_{s(b)}$ and their corresponding borders are $b_{\ (l)}$, $b_{(r),}$ $b_{\ (t)}, b_{\ (b)}$, respectively. We use a size parameter s to define four regions $r_{s(r)}$, $r_{s(l),} r_{s(t)}, r_{s(b)}$ as following.

$$r_{s(l)} = \{(x, y) | 1 \leq x \leq \frac{w}{s},\ 1 \leq y \leq h\}$$

$$r_{s(r)} = \{(x, y) | \frac{w(s-1)}{s} \leq x \leq w,\ 1 \leq y \leq h\}$$

$$r_{s(t)} = \{(x, y) | 1 \leq x \leq w,\ 1 \leq y \leq \frac{h}{s}\}$$

$$r_{s(b)} = \{(x, y) | 1 \leq x \leq w, \frac{h(s-1)}{s} \leq y \leq h\}$$

We define their corresponding borders b $_{(l)}$, b$_{(r),}$ b $_{(t)}$, b $_{(b)}$ as following:

$$b_{(l)} = \{(x, y) | 1 = x,\ 1 \leq y \leq h\}\ , \tag{3}$$

$$b_{(r)} = \left\{(x, y) | x = w,\ 1 \leq y \leq h\right\}, \tag{4}$$

$$b_{(t)} = \left\{(x, y) | 1 \leq x \leq w,\ 1 = y\right\}, \tag{5}$$

$$b_{(b)} = \{(x, y) | 1 \leq x \leq w, y = h\}. \tag{6}$$

The definition of 4 regions and their corresponding borders is shown in Figure 2. The left region and left border are shown and other regions can be shown similarly. Then, we examine four sides, respectively. We take left 1/s region $r_{s(l)}$ as an example. If there are foreground seeds in $r_{s(l)}$, retain the original edge. If there are no foreground seeds in $r_{s(l)}$, label the background seeds to the left boundary $b_{(l)}$, which is that region corresponding edge. Similarly, we examine the other three regions $r_{s(r)}, r_{s(t)},\ r_{s(b)}$.

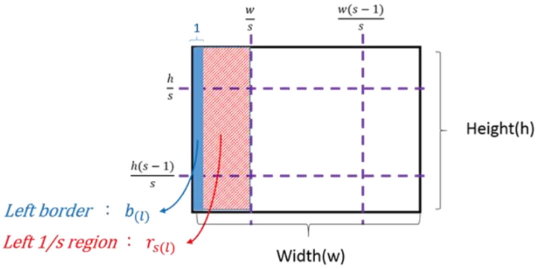


Figure 2.The definition of 4 regions and their corresponding borders.

After four sides examination, the extracted foreground still has some details which need to be improved. Thus, we apply Canny edge detection [8] to improve the segmentation. We find that the pixel in the edge image always belongs to the background in the saliency image. The green area is the background and the red area is the foreground. We use $(F_T - B_T)$/t to determine whether label the edge or not, where t is a parameter. The decision rule is as following：

For each pixel x_i in the edge image by Canny edge detection and its neighboring pixel$x_{i'}$. S(x_i) is the saliency value of pixel x_i, and S($x_{i'}$) is the saliency value of pixel $x_{i'}$.
if S($x_{i'}$)−S(x_i) is larger than $(F_T - B_T)$/t or S($x_{i'}$)−S(x_i) is equal to $(F_T - B_T)$/t, $x_{i'}$ is labelled *as* the foreground seed and x_i is la*bel*led *as* the background seed; otherwise, do nothing.

After automatically seeds labelling, we carry out the method of nonparametric higher-order learning for segmentation [3].

IV. Experiment Results and Analysis

We use our method to label the seeds efficiently and automatically. We compare with the method of nonparametric higher-order learning for interactive segmentation [3], which is greatly affected by users. The results of our method are very close to the results of nonparametric higher-order learning for interactive segmentation [3]. In addition, we also compare our method with the method [10], which also segment the image based on saliency. The parameters in our proposed method are presented in Table 1. Figure 3 and Figure 4 show the segmentation results of the image. We compare the result of our method with the result of Kim's method (supervised method) [3] first. According to Figure 3, it shows that our result is good and very close to the result of Kim's method (supervised method), which is a greatly affected by the users, while our method is unaffected by the users. Besides, we also compare the result of our method with the result of the method of Donoser et. al. [10]. According to Figure 4, the result of our method is better than that of the method [10], which is also an unsupervised segmentation method based on saliency because our method can label 2 bulls (foreground) only, but without the line between the grass and the lake.

V. Conclusion

In this paper, we proposed an unsupervised method based on saliency detection and edge detection to automatically and efficiently label the seeds. It is convenient because it can segment the foreground and background automatically and it doesn't need user interaction, which is quite important to segment a large data base of images.

References

[1] L. Grady, "Random walks for image segmentation," IEEE Transactions on Pattern Analysis and Machine Intelligence, 28(11):1768–1783, 2006.

[2] P. Kohli, L. Ladicky, and P. H. S. Torr," Robust higher order potentials for enforcing label consistency," Int J Computer Vision , 82:302–324, 2009.

[3] T.H. Kim, K.M. Lee, S.U. Lee, "Nonparametric higher-order learning for interactive segmentation," in: Computer Vision and Pattern Recognition (CVPR), 2010.

[4] H. Jiang, J. Wang, Z. Yuan, T. Liu, and N. Zheng," Automatic salient object segmentation based on context and shape prior," In British Machine Vision Conference, , 2011, pages 110.1–110.12 .

[5] Stas Goferman, Lihi Zelnik-Manor, and Ayellet Tal," Context-aware saliency detection," In CVPR, 2010, pages 2376–2383.

[6] K. Y. Chang, T. L. Liu, H. T. Chen, and S. H. Lai, "Fusing Generic Objectness and Visual Saliency for Salient Object Detection," in IEEE International Conference on Computer Vision, 2011, pp. 914-921.

[7] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," In CVPR, 2013.

[8] Canny, J., A. "A Computational Approach To Edge Detection" IEEE Transactions on pattern analysis and machine intelligence, 1986, VOL. PAMI-8, NO. 6, Page 679 – 698.

[9] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "Slic superpixels," 2010, EPFL Technical Report no. 149300.

[10] M. Donoser, M. Urschler, M. Hirzer, and H. Bishof, "Saliency driven total variational segmentation," in Proc. of the IEEE Int'l Conf. Computer Vision (ICCV'09), 2009.

[11] Baradez, M.O., McGuckin, C.P., Forraz, N., Pettengell, R., Hoppe, A.: 'Robust and automated unimodal histogram thresholding and potential applications," Pattern Recognit., 2004, 37, (6), pp. 1131–1148

[12] A. K. Jain and R. C. Dubes, Algorithms for Clustering Data. Prentice Hall, 1988..

[13] F. Kurugollu, B. Sankur, and A. Harmanci, "Color image segmentation using histogram multithresholding and fusion," Image and Vison Computing," 2001, 19(13):915–928.

[14] Y. Deng, B. S. Manjunath, "Unsupervised segmentation of color-texture regions," IEEE Trans. on Pattern Analysis and Machine Intelligence, 2001, 23(8):800~810.

[15] Y. Deng, C. Kenney, M. S. Moore , B. S. Manjunath, "Peer group filtering and perceptual color image quantization," In: Proc. of the IEEE Int'l Symp. on Circuits and Systems, 1999. 21~24.

ISBN: 978-1-61208-452-7

FG/BG threshold finding	b=3
	f=100

Four sides examination	s=8
Combine edge/saliency detection	σ (Standard Deviation)： 3 double threshold： 0.1 and 0.2 t=3

Table 1. The parameters in our proposed method.

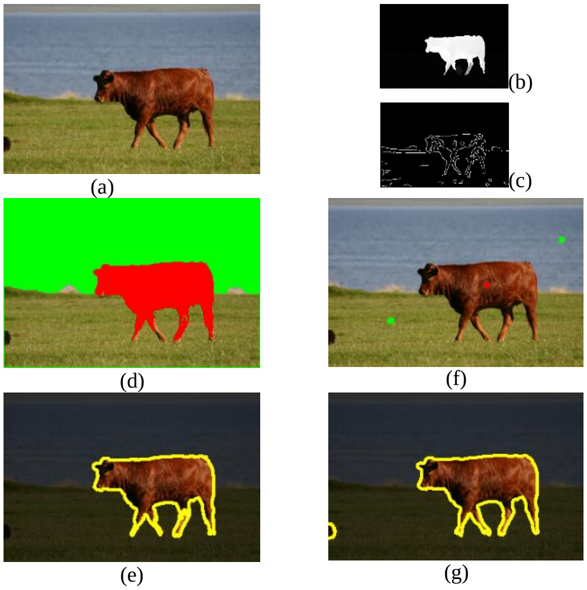


Figure 3. Segmentation result of the image. (a) Input image. (b) Result of saliency detection. (c) Result of edge detection. (d) Our scribble. (e) Our result. (f) Kim’s scribble. (g) Kim’s result.

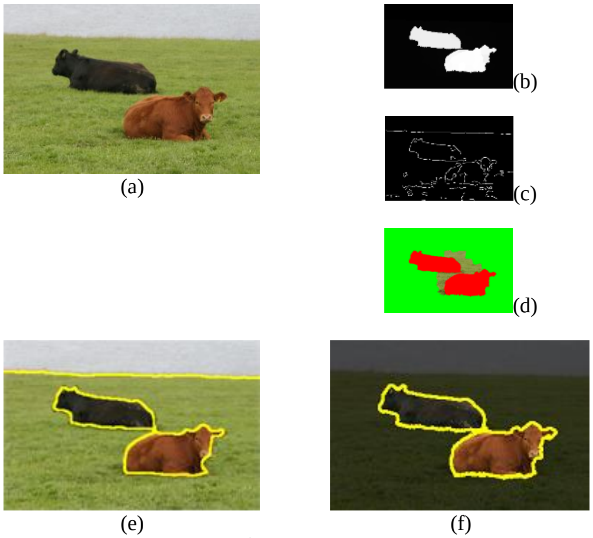


Figure 4. Segmentation result of the image. (a) Input image. (b) Result of saliency detection. (c) Result of edge detection. (d) Our scribble. (e) the result of Donoser et al. [10] (f) Our result.

Supervised Approach for Indication of Contrast Enhancement in Application of Image Segmentation

Gabriel F. C. Campos* , Rodrigo A. Igawa* , José L. Seixas Jr.* , Alex M. G. de Almeida* ,
Rodrigo Capobianco Guido† and Sylvio Barbon Jr.*

* Department of Computer Science, State University of Londrina, Brazil
† Department of Computer Science and Statistics, São Paulo State University, Brazil
Email: camposg@uel.br, igawa.rodrigo@gmail.com, jlseixasjr@gmail.com,
alex.marino@fatecourinhos.edu.br, guido@ieee.org, barbon@uel.br

Abstract—**Segmentation methods need a satisfactory input image with a good contrast between Region of Interest and background to provide a high accuracy result. Thus, the application of contrast enhancement before the segmentation is a usual practice. This paper presents a supervised approach to determine if an input image is satisfactory for a specific segmentation approach by using Feature Extraction, Feature Selection and Machine Learning. Experimental results showed the proposed approach ability to indicate the need of contrast enhancement in different segmentation problems with 94 percent accuracy.**

***Keywords*–Contrast; Image Segmentation; Machine Learning.**

I. Introduction

Image segmentation is one of the most difficult tasks in image processing [1] [2] [3]. In order to provide high accuracy results, segmentation methods need a satisfactory input image with good contrast between Region of Interest (ROI) and background. Due to this fact, application of contrast enhancement before segmentation is an usual practice as shown in [4] [5] [6] [7] [8].

Some contrast enhancement algorithms, such as CLAHE [9], rely on input parameters to be performed. This aspect of contrast enhancement implies that an individual configuration for each image is necessary. A standard configuration for the whole image dataset is impractical in a real application.

Undesirable results, such as noise formation, are obtained in cases where contrast enhancement is used with wrong parameters or in an originally good contrast image, especially when using a simple and general contrast enhancement algorithm like Histogram Equalization (HE). Furthermore, execution time may decrease considerably by removing a useless contrast enhancement step.

The features of an image provide useful information for automatic classification. In this paper, we explore several features, like histogram-based [10], gray-level co-occurrence matrix [11] [12] and Fast Fourier Transform (FFT) [13] [14], which can retrieve contrast and texture information from an image. Our proposal is a supervised approach with image features as input to classify between insufficient and sufficient images contrast, making it possible to decide if a contrast enhancement is necessary before the segmentation step, regardless the problem. The appropriate contrast enhancement step for images with insufficient contrast was not addressed in this paper.

Works related to our proposal are presented in Section II, while a detailed description of our proposed approach is presented in Section III. In Section IV, materials, methods and experiments used to validate our proposed approach are described and in Section V, we show results and discussion of this work. Conclusions are presented in Section VI.

II. Related Work

A good contrast between ROI and background in an input image is essential to provide high accuracy segmentation results. Due to this fact, application of contrast enhancement before segmentation is a usual practice.

In [15], an enhancement approach is presented to address the limitations of medical thermal images such as low contrast, low signal-to-noise ratio, and absence of clear edges. Despite these limitations which usually make the segmentation process difficult, the proposed approach using image enhancement were able to segment the images with an average accuracy of 98%. Similarly, [16] shows that pre-processing can have positive impacts on mammographic segmentation since it improves the contrast of tissue structures in uncompressed breast peripheral areas. Those recent works, and others like [6] and [8], indicate that contrast enhancement before segmentation can be useful to improve segmentation accuracy in gray-scale images.

The CLAHE contrast enhancement algorithm, which relies on input parameters to be performed, is used to improve segmentation as well. In [4], CLAHE was used to successfully improve the accuracy of a fruit segmentation approach. Even if color images were acquired, CLAHE was applied to the Intensity channel (gray-scale) and the enhanced image was segmented by the Hough algorithm. Fixed parameters were used for CLAHE in this approach.

CLAHE was used again in [5], with the goal of improving the segmentation in an intelligent iris recognition system for eye images. Regarding the two main CLAHE parameters, the clip limit parameter was dynamically chosen by the technique proposed in [17], while the sub-region size was fixed to 8x8.

The use of fixed parameters to improve a set of images is not the better solution, since some images may need more enhancement than others. It is possible as well that some images do not require any enhancement (which can be a problem even to contrast enhancement algorithms that do not need parameters). In a real world dataset, in which thousands of images will be segmented, enhancing images that do not require any enhancement may represent a huge waste of time depending on the algorithm used. An unnecessary enhancement can also create noise and decrease segmentation accuracy. Thus, it is important to know which images must be enhanced before segmentation.

Some applications use the information contained in multiple channels of an image to perform segmentation. This information is usually related to color. Segmentation and enhancement of color images are not handled in the actual stage of our work, but the idea of enhancing images to improve its segmentation can be applied to color images as well. It can be seen in [7] and [18].

Since features can provide useful information of an image, they may be used for automatic classification. Our proposed approach uses a supervised classifier based on Machine Learning to classify images considering its features.

Machine Learning for Image Classification is widely used in Medical applications [19] [20] [21] [22] [23]. In [24], a survey about Medical image analysis with artificial neural networks is presented. It shows that, besides segmentation, Machine Learning can be useful to classify images and ROI's, providing computer-aided detection and diagnosis.

Machine Learning is also used for aerial and satellite image classification [25] [26] [27], image classification in agriculture applications [28], in astronomical applications [29], image classification in palynology [30] and many other image applications.

Still, contrast enhancement before segmentation can improve accuracy. It may be useful to know which images must be enhanced before segmentation. By extracting the right features, it is possible to classify images using machine learning, which motivates our proposal: a supervised approach to indicate the need of contrast enhancement in applications of image segmentation.

III. Proposed Approach

Inadequate usage of contrast enhancement leads to wrong segmentation and misunderstandings concerning final results, since some features like objects intensity, objects size, area of image occupied by objects and number of different objects are relevant while applying an automatic segmentation approach.

We propose, as shown in Figure 1, an approach to build a model capable of identifying images with insufficient contrast for a specific segmentation task. The built model is created based on feature vectors extracted from a small set of image examples, Image Subset, and it is improved using feature subset selection, considering a supervised labeling process.

A specialist labels each image of the Image Subset as insufficient or sufficient, and this, combined with the extracted features, provide a training set. After performing a feature subset selection, the training set can be used as input to the Classifier Training step, as shown in Figure 1. Labels are applied by visual evaluation from original image and segmentation result. The Training Set is composed of a subset of features that optimally represent the focused segmentation related to the contrast enhancement applied.

In the last step, Supervised Classifier Training, the model for image classification between insufficient or sufficient contrast is created.

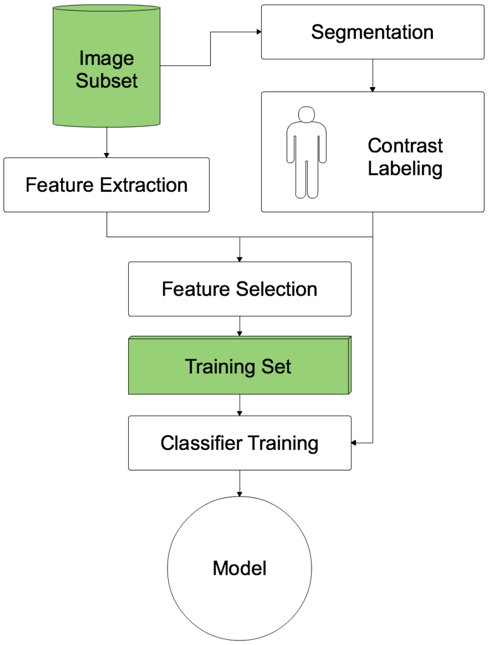


Figure 1. Proposed Approach

A. Image Features

A histogram is a statistical tool that can be used in contrast quality assurance and can represent the mean luminosity of an image. Considering this, we use different metrics from the histogram associated with Texture (Gray-tone spatial dependencies and Spectral Analysis) with the purpose of covering distinct image applications.

Texture is an important characteristic used to identify objects or regions of interest in images. Texture features based on gray-tone spatial dependencies are easily computable and have a general applicability for a wide variety of image-classification applications [11]. Gray-level co-occurrence matrices indicate how often a pixel with gray-tone value i occurs horizontally adjacent to a pixel with value j [12]. On the other hand, texture features based on Spectral Analysis can detect global periodicity on images by finding narrow peaks of high energy in frequency (spectrum) domain and Fourier transform is a common spectral method [13] [14].

We selected several features based on Histogram, Gray-tone spatial dependencies and Image Fourier Domain (shown in Table I). In Figure 2, it is possible to see features P12, P13, P14, P15, P16 and P17 in comparison to a standard histogram, where the x-axis represents gray tones going from 0 to 255 and the y-axis represents the frequency of each tone in an image.

B. Correlation-based Feature Subset Selection

A central problem in machine learning is identifying a representative set of features that best represents a model, increasing precision and reducing dimension. In this paper, we addressed this problem through a correlation based approach, where the main hypothesis is that good feature sets contain features that are highly correlated with the insufficient or sufficient label, yet uncorrelated with each other [31].

Concretely, a correlation-based approach is an algorithm which evaluates a great number of features subsets in order to

TABLE I. IMAGE FEATURES: HISTOGRAM, TEXTURE AND SPECTRAL

ID	Description
P1	Entropy of original grayscale image
P2	Entropy of gray-level co-occurrence matrix
P3	Inertia of gray-level co-occurrence matrix
P4	Energy of gray-level co-occurrence matrix
P5	Correlation of gray-level co-occurrence matrix
P6	Homogeneity of gray-level co-occurrence matrix
P7	Entropy of FFT
P8	Energy of FFT
P9	Inertia of FFT
P10	Homogeneity of FFT
P11	Image resolution
P12	Amount of non-zeros index
P13	Amount of non-zeros groups
P14	Largest length group
P15	Smallest length group
P16	Peak of Largest length group
P17	Peak of Smallest length group
P18	Amplitude of mean
P19	Amplitude of median
P20	Histogram Variance
P21	Histogram Standard Deviation

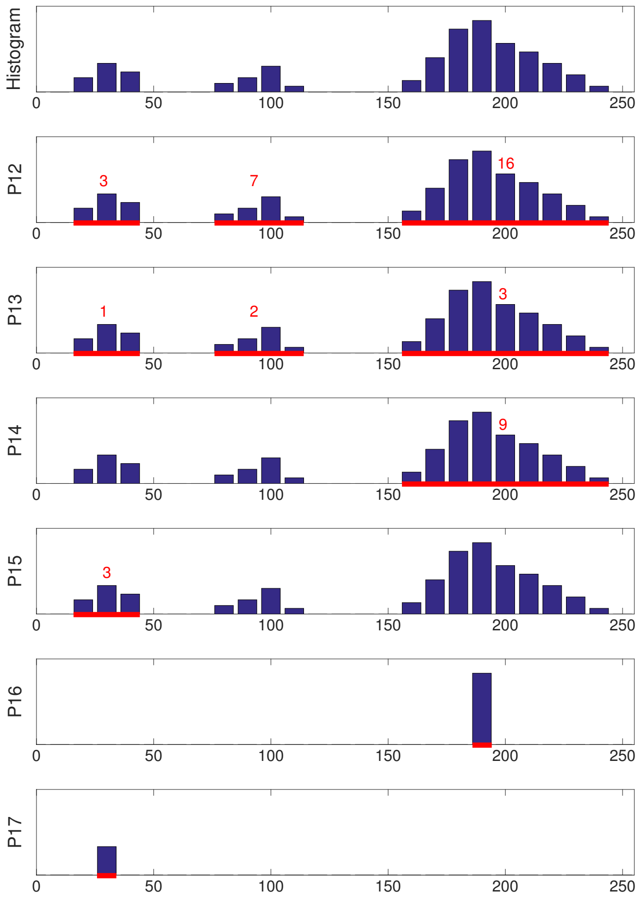


Figure 2. Histogram Based Features

obtain a better set of features than the current one. In order to do so, a correlation-based algorithm is initialized with an empty set of features. Then, with each round, a new feature is added to the set and measures like entropy, relief or merit are used to evaluate how suitable that subset of features has become. The most usual way to add features to the subset is by using a best first search in a feature space [31].

C. *Classification*

In order to have a suitable approach, which can be used for different segmentation problems and contrast enhancers, a supervised classifier based on Machine Learning is required. This way, it is possible to inform classifiers about what feature can improve ROI segmentation and when the contrast between the ROI and the background is good enough to a specific segmentation process.

Another important characteristic of several Machine Learning approaches, as Artificial Neural Networks (ANN), is online learning [32]. The online learning capacity means that non-stationary processes, which this might well be, can be modelled dynamically based on new image samples. Furthermore, the generalization and scaling feature leads to a model based on fewer samples [33]. In this paper, ANN were chosen as Machine Learning classifiers, since they are widely used in classification problems [24] [34] [35] [36] [37].

IV. Experimental Settings

Two datasets were used in experiments. The first one (Dataset I) was composed of medical images, where the ROI was regarding a wound region. The second dataset (Dataset II) was composed of pork image samples where the ROI was the intramuscular fat (marbling).

Note that the datasets represent different real world problems, specifically regarding ROI, where size, color and contrast with the background is really different between both datasets.

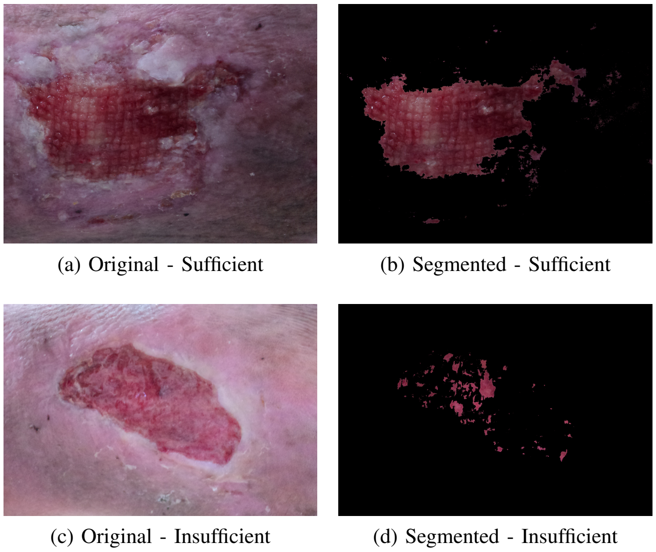
Three experiments were conducted in order to validate our proposed approach.The first experiment was performed using the medical dataset, the second was performed using the pork dataset and the third experiment was performed by combining both datasets in order to verify the robustness of the model to handle two different image scenarios.

The medical images dataset was composed of 100 color images. All files were in Portable Network Graphics format (PNG) and an example image can be seen in Figure 3a. The complete information about the image acquisition for the medical images dataset can be found in [38].

The pork images dataset was composed of 300 grayscale images containing meat samples (the background was already removed). All files were PNG, as well and an example image can be seen in Figure 4a. The pork images dataset was acquired using a digital single-lens reflex camera and a tripod that supported the device at 37cm above the sample. The camera was configured with automatic settings and had a 16.2 megapixels image sensor and high quality lens, which was optimally engineered to gather more light.

All images were segmented by thresholding, where threshold value was found by max entropy algorithm [39]. The medical images dataset was segmented in the same way as the pork images dataset, but, in order to obtain a better visualization in the labeling process, the original image colors were applied instead of saturation values used in segmentation.

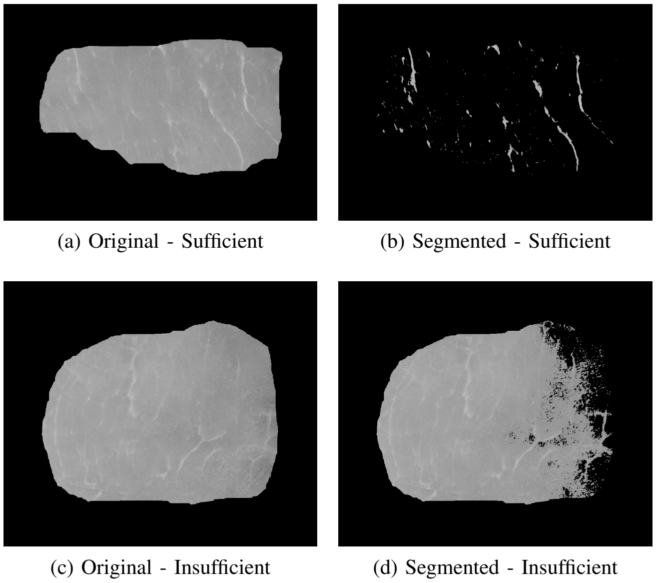
Figure 3 shows the segmentation of images from the medical dataset. Figures 3a and 3b represent samples labeled as ’sufficient contrast’ while Figures 3c and 3d represent samples labeled as ’insufficient contrast’.


(a) Original - Sufficient (b) Segmented - Sufficient

(c) Original - Insufficient (d) Segmented - Insufficient

Figure 3. Medical dataset - Segmentation and labeling

Figure 4 shows the segmentation of images from the pork dataset. Figures 4a and 4b represent samples labeled as ’sufficient contrast’ while Figures 4c and 4d represent samples labeled as ’insufficient contrast’.


(a) Original - Sufficient (b) Segmented - Sufficient

(c) Original - Insufficient (d) Segmented - Insufficient

Figure 4. Pork dataset - Segmentation and labeling

As it can be seen in Figure 3 and Figure 4, the labeling is simple and easy for such images, so only one specialist performed the labeling process. Every image that generated any kind of doubt in the evaluation by the specialist was removed from the dataset.

Before the labeling process, random samples were removed to balance both datasets. Thus, we obtained Dataset I Balanced (Dataset I BA) and Dataset I Imbalanced (Dataset I IM), as well, Dataset II Balanced (Dataset II BA) and Dataset II Imbalanced (Dataset II IM). An extra dataset was created by combining both balanced datasets (Dataset I BA and II BA).

In our experiment, the max entropy algorithm was used to find a threshold value for image segmentation. It is a simple and fast method, being one of the best solutions to segment meat marbling [39] [40], which is one of our datasets.

In order to evaluate the generalization power of ANN, we experimented with two different networks: Multilayer Perceptron (MLP) and Radial Basis Function (RBF) [19] [41] [42] [43] [44]. We evaluated approximately 22 different MLP and RBF architectures.

Regarding computational complexity, the MLP is $O(n2)$ to train and $O(n)$ to execute while the RBF is $O(n)$ to both train and execute [19] [45]. Fortunately, the specialist only needs to label a small subset of images in order to build the training set, which makes the model viable to be used in real-world computer vision problems with huge datasets.

Towards MLPs, the following settings values were used: learning rate 0.3, momentum 0.2, number of epochs 2500. For MLP, we experimented with different numbers of perceptrons on the hidden layers, ranging from 1 to 22. Settings values towards RBF, on the other hand, were: number of basis functions ranging from 1 to 22, number of iterations on logistic regression until convergence and seeds on k-means as defaults.

The results are mostly discussed in terms of accuracy. However, in special cases, detailed results can be shown in confusion matrices. As it is known in many cases where supervised learning is performed, a confusion matrix is represented by a simple two dimensional matrix. Rows represent the actual classes of each instance, in our case contrast sample labeled. Columns represent the predicted classes, for us: contrast sufficient or insufficient based on the supervised model. This way, ideal results correspond to large numbers on the main diagonal and smaller numbers, hopefully zero, for entries off the main diagonal.

V. Results and Discussion

An overview of the results is shown in Table II. The results of the most accurate architectures are shown for each classifier in each dataset. The results regarding feature selection are also presented.

We first discuss feature selection relevance. As it is possible to note, feature selection matters not only to increase performance by dimensional reduction, but also to help with accuracy. This importance is shown in Table II, where column AF Acc (All Features Accuracy) always presented lower results than column SF Acc (Selected Features Accuracy). In general, as shown in the lower part of Table II, feature selection added an average 6.04% accuracy. There is only one case where using feature selection did not improve classifier accuracy, which is RBF in Dataset II IM. Details considering this special case are properly discussed in Section V-B.

Considering classifiers, it is also possible to realize that the best results from MLP were always higher than the best results from RBF. Details concerning accuracies are shown in Figure 5 and discussed in Section V-B.

A. Feature Selection

Towards selected features, Table III shows results considering all datasets. As it can be seen in the second column

TABLE II. ACCURACIES OF CLASSIFIERS WITH ALL AND SUBSET FEATURES

Dataset	ANN	AF Acc	SF Acc	Diff.
I IM	MLP	87.21%	88.72%	1.51%
	RBF	81.11%	82.20%	1.09%
I BA	MLP	86.13%	91.08%	4.95%
	RBF	67.32%	80.19%	12.87%
I & II BA	MLP	87.06%	89.55%	2.49%
	RBF	72.13%	82.58%	10.45%
II IM	MLP	83.66%	84.66%	1.00%
	RBF	83.33%	83.33%	0.00%
II BA	MLP	88.00%	94.00%	6.00%
	RBF	70.00%	90.00%	20.00%
			Mean Diff.	6.04%

(quantity of features selected), dimensionality reduction was significant. In the particular case of Dataset II IM, it was possible to reduce the number of features from 22 to just 4. In another case, as seen for Dataset I BA, dimensionality reduction was less expressive, but still, reduced from 22 to just 9 features.

In summary, the features selected for each dataset varied. No feature was present in all cases after feature selection. The most used feature was P12 but this feature was not present in Dataset II BA feature selection.

TABLE III. SUB-SELECTED FEATURES IN EACH EXPERIMENT

DataSet	Quantity	Features Subset
I IM	7	P12, P14, P17, P3, P4, P8 and P7
I BA	9	P1, P12, P14, P17, P2, P3, P4, P6 and P8
I & II BA	5	P1, P12, P13, P2 and P3
II IM	4	P12, P13, P17 and P2
II BA	6	P11, P16, P17, P18, P21 and P9

B. *Classifiers*

Regarding classifiers, Figure 5 shows box-plots for both MLP and RBF on each dataset when using feature selection. MLP, plotted as blue box-plots, shows much greater accuracy than RBF in all five datasets. Although MLP presented lower outliers in some cases, they were still better than the RBF results. This fact is notable on Dataset I IM and Dataset I & II BA where the lowest results from MLP are still higher than the median results from RBF. Another outstanding result achieved by MLP on experiments is the difference from third quartile and the first quartile in all datasets. In practice, this is shown on the smaller box size from MLP plots which results in a very stable classifier on experiments.

Considering balanced and imbalanced instances, it is also possible to realize that MLP achieved higher results when the dataset is balanced, as seen from Dataset I BA to Dataset I IM and from Dataset II BA to Dataset II IM. However, this is not the case for RBF. Actually, on Dataset I IM, RBF achieved better results than in Dataset I BA, where instances are balanced.

A third situation is shown by results on Dataset I & II BA, a situation where both scenarios are tested at once. This time, MLP also performed better than RBF regarding stability and higher accuracy results. Considering stability, MLP presented only one outlier, while RBF presented several. Considering higher accuracy results, the entire box from MLP is plotted above the best results from RBF.

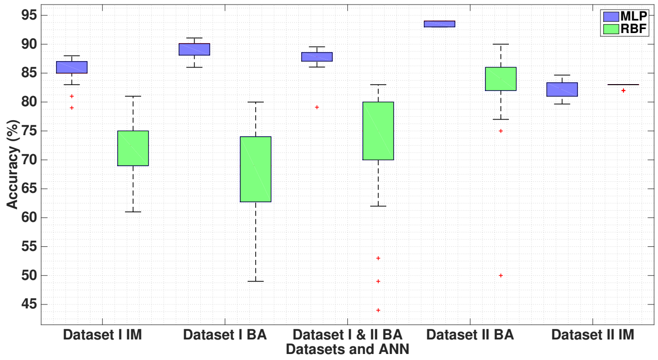


Figure 5. Boxplot of MLP and RBF accuracies

TABLE IV. CONFUSION MATRIX OF DATASET II IM PERFORMED BY RBF

		Predicted		
		Sufficient	Insufficient	Total
Actual	Sufficient	250	0	250
	Insufficient	50	0	50
	Total	300	0	300

TABLE V. CONFUSION MATRIX OF DATASET II BA PERFORMED BY MLP

		Predicted		
		Sufficient	Insufficient	Total
Actual	Sufficient	46	4	50
	Insufficient	2	48	50
	Total	48	52	100

A special case is shown in the case of Dataset II IM, where RBF seemed stable by presenting a very small box. However, it is not a true sign of success. As shown in Table IV, the results imply that, actually, the classifier presented a very high rate of error. Therefore, in different datasets and different ways of evaluation, MLP presented more desirable results than RBF. In order to show that results in other cases were satisfactory in terms of generalization, we show the prediction results to another classifier on Dataset II BA. Table V shows predictions with settings which achieved 94% accuracy to classify instances.

VI. Conclusion

In this paper, we proposed a supervised approach able to indicate the need of contrast enhancement in images, specifically before segmentation process. Experimental results showed high accuracy when dealing with two different datasets, especially when using feature selection and MLP classifier. As well, it exposed some features that best represent contrast in digital images.

The proposed approach can be used in computer vision systems that can afford a segmentation step, avoiding undesirable noise and wasted time by an incorrect or useless application of a contrast enhancement method.

Addressing feature discussions, dimensional reduction was an important issue on our approach in terms of performance. As discussed in Section V-A, in a particular case, feature selection enabled reduction from 22 to only 4 features.

ANN accuracy was also discussed and experiments showed that, mostly, MLP performed better than RBF in terms of maximum accuracy, outliers values and stability. Details concerning such issue were presented in Section V-B. Still, RBF presented a very peculiar case in which no generalization was performed correctly, as seen in Table IV.

In summary, supervised learning approach for indication of contrast enhancement in image segmentation was successfully achieved. Different datasets were tested and 94% accuracy was achieved in a case where different datasets were tested at the same time. The proposed approach performed well not only when the dataset consisted of a single scenario, but also when the dataset consisted of different image scenarios.

As future work, we will employ contrast enhancement approach in order to observe the behavior of the classifier and analyze if the model will be able to handle the adjusted contrast. We also intend to use color information as features besides segmenting the images with more complex algorithms, e. g. watershed and decision trees.

ACKNOWLEDGMENT

The authors would like to thank CAPES (Brazilian Agency) for financial support.

REFERENCES

[1] R. C. Gonzalez and R. E. Woods, Processamento Digital De Imagens. Addison Wesley Bra, 2010.

[2] T. Sag and M. Çunka, "Color image segmentation based on multi-objective artificial bee colony optimization," Applied Soft Computing, vol. 34, no. 0, 2015, pp. 389 – 401.

[3] W. Shi, R. Zou, F. Wang, and L. Su, "A new image segmentation method based on multifractal detrended moving average analysis," Physica A: Statistical Mechanics and its Applications, vol. 432, no. 0, 2015, pp. 197 – 205.

[4] E. A. Murillo-Bracamontes et al., "Implementation of hough transform for fruit image segmentation," Procedia Engineering, vol. 35, no. 0, 2012, pp. 230 – 239, international Meeting of Electrical Engineering Research 2012.

[5] A. F. M. Raffei, H. Asmuni, R. Hassan, and R. M. Othman, "A low lighting or contrast ratio visible iris recognition using iso-contrast limited adaptive histogram equalization," Knowledge-Based Systems, vol. 74, no. 0, 2015, pp. 40 – 48.

[6] L.-C. Chen, C.-H. Chien, and X.-L. Nguyen, "An effective image segmentation method for noisy low-contrast unbalanced background in mura defects using balanced discrete-cosine-transfer (bdct)," Precision Engineering, vol. 37, no. 2, 2013, pp. 336 – 344.

[7] G. Schaefer, M. I. Rajab, M. E. Celebi, and H. Iyatomi, "Colour and contrast enhancement for improved skin lesion segmentation," Computerized Medical Imaging and Graphics, vol. 35, no. 2, 2011, pp. 99 – 104, advances in Skin Cancer Image Analysis.

[8] N. Al-Najdawi, M. Biltawi, and S. Tedmori, "Mammogram image visual enhancement, mass segmentation and classification," Applied Soft Computing, vol. 35, no. 0, 2015, pp. 175 – 185.

[9] K. Zuiderveld, "Graphics gems iv," P. S. Heckbert, Ed. San Diego, CA, USA: Academic Press Professional, Inc., 1994, ch. Contrast Limited Adaptive Histogram Equalization, pp. 474–485.

[10] G. Balaji, T. Subashini, and N. Chidambaram, "Automatic classification of cardiac views in echocardiogram using histogram and statistical features," Procedia Computer Science, vol. 46, no. 0, 2015, pp. 1569 – 1576, proceedings of the International Conference on Information and Communication Technologies, {ICICT} 2014, 3-5 December 2014 at Bolgatty Palace amp; Island Resort, Kochi, India.

[11] R. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," Systems, Man and Cybernetics, IEEE Transactions on, vol. SMC-3, no. 6, Nov 1973, pp. 610–621.

[12] S. Chowdhury, B. Verma, and D. Stockwell, "A novel texture feature based multiple classifier technique for roadside vegetation classification," Expert Systems with Applications, vol. 42, no. 12, 2015, pp. 5047 – 5055.

[13] H.-K. Shen, P.-H. Chen, and L.-M. Chang, "Automated steel bridge coating rust defect recognition method based on color and texture feature," Automation in Construction, vol. 31, no. 0, 2013, pp. 338 – 356.

[14] M. Nixon and A. Aguado, Feature Extraction and Image Processing, ser. Electronics & Electrical. Newnes, 2002.

[15] S. Suganthi and S. Ramakrishnan, "Anisotropic diffusion filter based edge enhancement for segmentation of breast thermogram using level sets," Biomedical Signal Processing and Control, vol. 10, 2014, pp. 128 – 136.

[16] W. He, P. Hogg, A. Juette, E. R. Denton, and R. Zwiggelaar, "Breast image pre-processing for mammographic tissue segmentation," Computers in Biology and Medicine, vol. 67, 2015, pp. 61 – 73.

[17] B. S. Min, D. K. Lim, S. J. Kim, and J. H. Lee, "A novel method of determining parameters of clahe based on image entropy," International Journal of Software Engineering and Its Applications, vol. 7, no. 5, 2013, pp. 113–120.

[18] A. Aimi Salihah, M. Mashor, N. Harun, A. Abdullah, and H. Rosline, "Improving colour image segmentation on acute myelogenous leukaemia images using contrast enhancement techniques," in Biomedical Engineering and Sciences (IECBES), 2010 IEEE EMBS Conference on, Nov 2010, pp. 246–251.

[19] J. Seixas, S. Barbon, and R. Gomes Mantovani, "Pattern recognition of lower member skin ulcers in medical images with machine learning algorithms," in Computer-Based Medical Systems (CBMS), 2015 IEEE 28th International Symposium on, June 2015, pp. 50–53.

[20] B. Ashinsky et al., "Machine learning classification of oarsi-scored human articular cartilage using magnetic resonance imaging," Osteoarthritis and Cartilage, vol. 23, no. 10, 2015, pp. 1704 – 1712.

[21] A. Gertych et al., "Machine learning approaches to analyze histological images of tissues from radical prostatectomies," Computerized Medical Imaging and Graphics, vol. 46, Part 2, 2015, pp. 197 – 208, information Technologies in Biomedicine.

[22] S. Yu, K. K. Tan, B. L. Sng, S. Li, and A. T. H. Sia, "Lumbar ultrasound image feature extraction and classification with support vector machine," Ultrasound in Medicine Biology, vol. 41, no. 10, 2015, pp. 2677 – 2689.

[23] H. Joutsijoki, M. Haponen, I. Baldin, J. Rasku, Y. Gizatdinova, M. Paci, J. Hyttinen, K. Aalto-Setala, and M. Juhola, "Histogram-based classification of ipsc colony images using machine learning methods," in Systems, Man and Cybernetics (SMC), 2014 IEEE International Conference on, Oct 2014, pp. 2611–2617.

[24] J. Jiang, P. Trundle, and J. Ren, "Medical image analysis with artificial neural networks," Computerized Medical Imaging and Graphics, vol. 34, no. 8, 2010, pp. 617 – 631.

[25] S. Manthira Moorthi, I. Misra, R. Kaur, N. Darji, and R. Ramakrishnan, "Kernel based learning approach for satellite image classification using support vector machine," in Recent Advances in Intelligent Computational Systems (RAICS), 2011 IEEE, Sept 2011, pp. 107–110.

[26] G. Yan and S. Fenzhen, "Study on machine learning classifications based on oli images," in Mechatronic Sciences, Electric Engineering and Computer (MEC), Proceedings 2013 International Conference on, Dec 2013, pp. 1472–1476.

[27] P. Jordhana and K. Soundararajan, "Kernel methods and machine learning techniques for man-made object classification in sar images," in Information Communication and Embedded Systems (ICICES), 2014 International Conference on, Feb 2014, pp. 1–6.

[28] S. Mulyono, T. Pianto, M. Fanany, and T. Basaruddin, "An ensemble incremental approach of extreme learning machine (elm) for paddy growth stages classification using modis remote sensing images," in Advanced Computer Science and Information Systems (ICACSIS), 2013 International Conference on, Sept 2013, pp. 309–314.

[29] M. Abd Elfattah, N. Elbendary, H. Elminir, M. Abu El-Soud, and A. Hassanien, "Galaxies image classification using empirical mode decomposition and machine learning techniques," in Engineering and Technology (ICET), 2014 International Conference on, April 2014, pp. 1–5.

[30] M. Popescu and L. Sasu, "Feature extraction, feature selection and machine learning for image classification: A case study," in Optimization

of Electrical and Electronic Equipment (OPTIM), 2014 International Conference on, May 2014, pp. 968–973.

[31] M. A. Hall, “Correlation-based feature selection for machine learning,” Tech. Rep., 1999.

[32] U. Seiffert, “Annieartificial neural network-based image encoder,” Neurocomputing, vol. 125, 2014, pp. 229–235.

[33] I. Basheer and M. Hajmeer, “Artificial neural networks: fundamentals, computing, design, and application,” Journal of microbiological methods, vol. 43, no. 1, 2000, pp. 3–31.

[34] O.-R. Manuel, M.-B. del Rosario, G. Eduardo, and V.-C. Rene, “Artificial neural networks modeling evolved genetically, a new approach applied in neutron spectrometry and dosimetry research areas,” in Electronics, Robotics and Automotive Mechanics Conference, 2008. CERMA ’08, Sept 2008, pp. 387–392.

[35] X. Li and X. Yu, “Influence of sample size on prediction of animal phenotype value using back-propagation artificial neural network with variable hidden neurons,” in Computational Intelligence and Software Engineering, 2009. CiSE 2009. International Conference on, Dec 2009, pp. 1–4.

[36] V. Shahpazov, V. Velev, and L. Doukovska, “Design and application of artificial neural networks for predicting the values of indexes on the bulgarian stock market,” in Signal Processing Symposium (SPS), 2013, June 2013, pp. 1–6.

[37] R. Venkatesan and B. Balamurugan, “A real-time hardware fault detector using an artificial neural network for distance protection,” Power Delivery, IEEE Transactions on, vol. 16, no. 1, Jan 2001, pp. 75–82.

[38] J. Seixas et al., “Color energy as a seed descriptor for image segmentation with region growing algorithms on skin wound images,” in e-Health Networking, Applications and Services (Healthcom), 2014 IEEE 16th International Conference on, Oct 2014, pp. 387–392.

[39] J. Kapur, P. Sahoo, and A. Wong, “A new method for gray-level picture thresholding using the entropy of the histogram,” Computer Vision, Graphics, and Image Processing, vol. 29, no. 3, 1985, pp. 273 – 285.

[40] K. Chen and C. Qin, “Segmentation of beef marbling based on vision threshold,” Computers and Electronics in Agriculture, vol. 62, no. 2, 2008, pp. 223–230.

[41] M. Bagheri, S. Mirbagheri, M. Ehteshami, and Z. Bagheri, “Modeling of a sequencing batch reactor treating municipal wastewater using multilayer perceptron and radial basis function artificial neural networks,” Process Safety and Environmental Protection, vol. 93, 2015, pp. 111 – 123.

[42] L. Cakir and N. Yilmaz, “Polynomials, radial basis functions and multilayer perceptron neural network methods in local geoid determination with gps/levelling,” Measurement, vol. 57, 2014, pp. 148 – 153.

[43] R. Velo, P. Lpez, and F. Maseda, “Wind speed estimation using multilayer perceptron,” Energy Conversion and Management, vol. 81, 2014, pp. 1 – 9.

[44] Z. Ramedani, M. Omid, A. Keyhani, S. Shamshirband, and B. Khoshnevisan, “Potential of radial basis function based support vector regression for global solar radiation prediction,” Renewable and Sustainable Energy Reviews, vol. 39, 2014, pp. 1005 – 1011.

[45] S. Haykin, Neural Networks: A Comprehensive Foundation. Prentice Hall, 1999.

A Crowd-sourced Method for Real-time Detection and Localization of Unexpected Events

Taishi Yamamoto

Graduate School of Information Science and Engineering
Ritsumeikan University
Shiga, Japan
Email: is0145rp@ed.ritsumei.ac.jp

Kenta Oku, Kyoji Kawagoe

College of Information Science and Engineering
Ritsumeikan University
Shiga, Japan
Email: {oku@fc, kawagoe@is}.ritsumei.ac.jp

Abstract—**In this paper, a Real-Time Smart and Quick Unexpected-Event Detect (RT-SQUED) method is proposed to detect and localize unexpected events, such as traffic accidents, in real-time, assuming crowd sourcing with smartphone devices. The authors previously proposed the SQUED, a crowd-sourced system for detection and localization of unexpected events from smartphone-sensor data. When SQUED users find an event, they point their smartphones toward a direction of the event. SQUED can detect the location and the time of the event using the smartphone-sensor data. However, the SQUED method is difficult to scale because of its computational cost problem. Therefore, it becomes difficult to detect events in real-time. Our new method, called RT-SQUED, solves this problem using concurrent two-phase processing. RT-SQUED is composed of rough and concrete processing phases. In this paper, effectiveness of RT-SQUED is also discussed.**

Keywords–event detection; smart phones; crowd-sourced system; global positioning system; sensor data.

I. Introduction

Recently, event detection using sensors, such as cameras or microphones, has attracted much attention owing to the spreading of such sensors. Moreover, many event-detection methods using setting-type and special-purpose sensors have been developed [1][2]. The popularization of smartphones is changing how events are detected, because a smartphone is equipped with many kinds of sensor devices, e.g., a camera, a microphone, a GPS, and an accelerometer. A special sensor does not need to be installed in a specific location in advance, because almost everyone possesses a smartphone. Although a smartphone sensor can be used for event detection, a crucial problem remains. Sensor data contains many errors because of its low measurement accuracy and the human's irregular hand movements.

We previously proposed a Smart and Quick Unexpected-Event Detector, SQUED, to detect unexpected events using built-in smartphone devices [3]. SQUED gathers data from the smartphones of people near the location of an unexpected event, such as a traffic accident. Both a location from the GPS and the direction from a geomagnetic sensor are used to identify the actual event location. The main points of SQUED are 1) crowd sourcing and 2) a novel event-detection method, allowing it to estimate an accurate location even from inaccurate data. SQUED can detect an event even if only two people are near the event location. Moreover, the more people who are around the location, the more accurately SQUED can detect the location, using our event-detection method. Finally, it is available in blind spots, which are the biggest problem for pre-installed event-detection equipment.

However, SQUED method has the following two problems. First, it is difficult to expand the detection range. The wider the detection range is, the more the event detection interval decreases. Second, it is difficult to detect multiple events occurred simultaneously. These problems are caused by high computational cost in detecting events.

In this paper, we propose an improved SQUED method, RT-SQUED, to be able to detect the event efficiently in a wider range. RT-SQUED performs two-phase processing, concurrently. In the first phase, RT-SQUED roughly divides its detection area into grid units for the event detection. Second, RT-SQUED counts the number of users whose eye's-view triangles overlapped in each grid. By these phases, it can recognize regions not have to search. Therefore, the computational cost can be reduced when the scaling of the detection range is allowed. In the second phase, RT-SQUED apply the original SQUED method to the grid in which the number of users whose eye's-view exceeds the threshold.

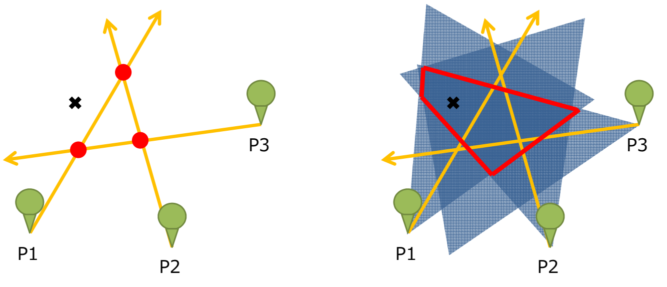
This paper is organized as follows. Section II discusses our previous work. The proposed event detection method is described in Section III. Section IV presents the results of evaluation and discussion of the proposed method. The related works are also explained in Section V. Finally, Section VI concludes this papers.

II. Previous Work

A. Basic concept

In Figure 1(a), the ”x” symbol indicates the actual location where an event occurred. Figure 1(a) shows three original directions obtained from the sensors of three pedestrians, P1, P2, and P3. Although each pedestrian tends to precisely point his/her smartphone toward the event location, an imprecise direction may be obtained from inaccurate sensor data, as shown in Figure 1(a). In Figure 1(a), the three direction lines do not intersect. Even for any pair of direction lines, the intersection is far from the actual event location.

The basic concept of our SQUED method is shown in Figure 1(b). In Figure 1(b), the eye's-view triangle introduced in our method is shown. The eye's-view triangle starts from the location of a pedestrian, with a pre-specified angle centering from the event direction estimated from the pedestrian's smartphone sensor. As in Figure 1(b), when an event occurs at location ”x”, a triangle for each pedestrian is constructed from



(a) Eyes direction (b) Eyes view

Figure 1. Eyes direction and eyes view

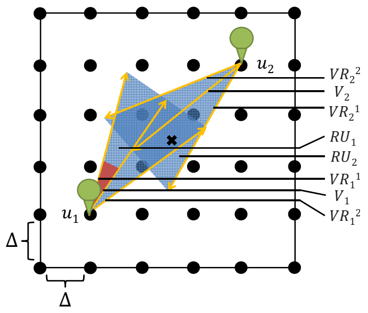


Figure 2. SQUED event detection method

the sensor data. Three triangles are checked in this case. The intersection of the three triangle regions is calculated. The red area, a parallelepiped represented by dashed lines, indicates the event location area, inside which the event likely occurred. With this introduction of an eye's-view triangle, the event-location area can be detected even when the observed sensor data contains noise, as in Figure 1(a). The more pedestrians who view an event and point a smartphone at it, the more precisely the location can be detected.

B. Definition

We define some symbols to describe our event detection method. R, a rectangular region, is pre-assigned as the event detection range. R is represented as a set of $\Delta \times n_x$ times $\Delta \times n_y$ grid points $\{P_{ij}\}$, $i = 1, .., n_x$, $j = 1, .., n_y$, which are disposed at equal intervals in both vertical and horizontal directions, as shown in Figure 2. Δ is the predefined interval between two adjacent grid points, in vertical and horizontal directions. The value of Δ is assumed to be having minimum accuracy on the detected event location. The left-bottom coordination point is $(B_x$,$B_y)$.

Suppose there are users $U = \{u_k\}$, $k = 1, .., N_u$ pointing their smartphones towards an event. For each user u_k, his/her location and eye's-view direction are defined as L_k and V_k, respectively. We introduce the eye's-view triangle represented as RU_k for a user U_k. RU_k is calculated from the origin point L_k and two lines from L_k whose angles are VR_k^1 and VR_k^2, respectively, where VR_k^1=V_k-α and VR_k^2=V_k+α, α is a pre-defined parameter. The length of the two lines are fixed as H.

C. Method

For a given TU_k, we calculate the number of overlapped users, $DU_{i,j}$ for each grid point $P_{i,j}$ as follows: $DU_{i,j} = |U^{'}_{i,j}|, whereU^{'}_{i,j} = \{u_l | P_{i,j} \in RU_l, u_l \in U\}$. We define the detected event region/location as E and the minimum number of overlapped users for event detection as MU. Then, the detected event region E is obtained as the following: $E = \{P_{i,j} | DU_{i,j} \geq MU\}$. The detected grid points of E are sorted with the value of $DU_{i,j}$ in descending order. Figure 2 shows an example of eye's-view triangles and an event-location region detected from the triangles. In this example, only two users found an event. This Figure has two eye's-view triangles. The intersection of the two regions is represented as a set of grid points, as shown in Figure 2. In our proposed method, we calculate the number of users whose eye's-view triangles overlap, rather than calculating a time-consuming intersection.

D. Problems of SQUED method

SQUED method has following problems: expansion of event detection range and detection of multiple events.

First, suppose that, we set Δ as 10 meters and the values of n_x and n_y as 1000. In this case, the number of grid points is 10000. For each of the grid points, SQUED calculate the number of overlapped users. As the number of users or $n_x \times n_y$ becomes large numbers, order of calculation will be larger. Therefore, it is difficult to expand the detection range.

Next, we assume that 10 events occur in a range of 5 km × 5 km at random. If the system can only search the range of 500 m × 500 m due to the above-mentioned problem, it is difficult to detect multiple events. In the case of expanding the event detection range, this problem becomes more serious.

Therefore, it is necessary to reduce the computational cost for event detection.

III. Proposed Event Detection Method

A. Basic concept

Figure 3 and Figure 4 show the rough processing phase, Figure 5 shows the concrete processing phase.

In the rough processing phase, it counts intersections between a rough grid dividing a detection range and a smallest rectangle surrounding the user triangle. In this example, the detection range is divided by four rough grids. Figure 3 shows the intersection detection in the left upper rough grid. Intersections in each rough grid are shown in Figure 4. Here, the rough grid which should detect in detail becomes only the left upper grid when the threshold is set to 2. Therefore, the concrete processing phase applies only to the left upper rough grid.

The concrete processing phase uses SQUED method which is mentioned above. In Figure 5, the part indicated as the red circle (the white circle in the gray scale) becomes the event detection point when the threshold is set to 2. Using this method, it can reduce the calculation cost.

B. Algorithms

Algorithm 1. EventDetection

```
EventDetection{
// RoughGridSize, ConcreteGridSize,
// DetectionRange, threshold are pre-defined.
// N-Interval, the number of concrete-phase
// processings in one rough-phase processing,
// is preseted.

```

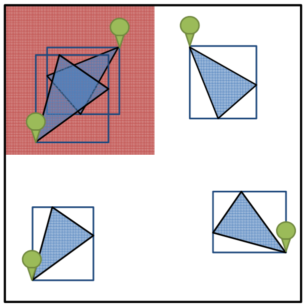


Figure 3. Rough processing phase

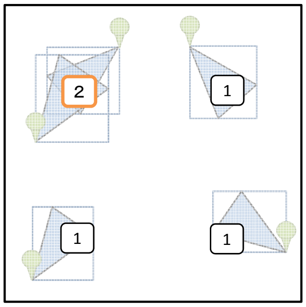


Figure 4. Rough processing phase 2

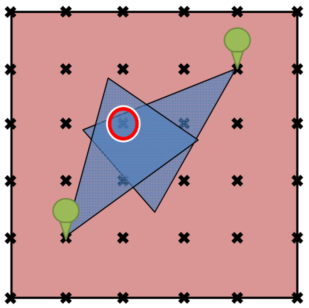


Figure 5. Concrete processing phase

```
//Initialization
 RoughGrid <- CreateGrid(RoughGridSize,
              DetectionRange)
 ConcreteGrid <- CreateGrid(ConcreteGridSize,
                 DetectionRange)
 //Main event detection process
 While(){
  If (Stop-condition is met) {Stop}
  Obtain all USER data from Database and
   set them to User.
  // An element of User is a set of three
  //  locations forming a triangle.
  If (Counter exceeds N-Interval) {
   DetectedRoughInfo <- RoughEventDetection
                        (RoughGrid, User)
   GridPointInfoUser <- GetConcreteGridPoint
           (ConcreteGrid, DetectedRoughInfo)
   Counter <- 0
  }
  DetectedResult <- ConcreteEventDetection
                    (GridPointInfoUser)
  Counter++
  ShowEventsDetected(DetectedResult)
 }
}
```

Algorithm 2. Rough & Concrete EventDetection

```
RoughEventDetection(RoughGrid, User) {
 DetectedRoughInfo <- {}
 For each grid in RoughGrid {
  For each user in User {
   MBR <- GetMinBoundingRect(user)
   // user’s MinimumBoundingRectangle is
   //  calculated.
   If (CheckIntersect(grid, MBR) is TRUE {
    // it is checked if the MBR intersects
    //  the grid or not.
    Add (grid, user) to DetectedRoughInfo
   }
  }
 }
 Return(DetectedRoughInfo)
}
ConcreteEventDetection(GridPointInfoUser) {
 DetectedResult <- {}
 For each grid point as gridPoint
  in GridPointInfoUser {
  // UCounter is a counter to count the number
  //  of users is included in the gridPoint.
  UCounter <- 0
  For each user in GridPointInfoUser {
   If (CheckInclusion(gridPoint, user)) is TRUE {
    // it is checked if the user is included
    //  the gridPoint or not.
    UCounter++
   }
  }
  If(UCounter >= threshold) {
   Add gridPoint to DetectedResult
  }
 }
 Return(DetectedResult)
}
```

In the Algorithm 1, the main process, called EventDetection, is described.

Each user has the positional information of the three points forming a triangle. Three points consists of the location of the user and two points making up the eye’s view is described in the previous section. A grid is a square composed of four coordinate grid points.

In EventDetection, first, two types of grid information are generated using RoughGrid and ConcreteGrid. CreateGrid is a function to create grids dividing from the DetectionRange at pre-defined size such as RoughGridSize. Then, RoughEventDetection, GetConcreteGridPoint and ConcreteEventDetection are iterated until the given stop condition is met, while RoughEventDetection and GetConcreteGridPoint are executed per preset N-interval. RoughEventDetection is a function to detect possible rough grids and users belonging to rough grids where a certain event may occur. GetConcreteGridPoint is a function to create a GridPointInfoUser, which is composed of a set of grid points and the users extracted for the grid point. The grid points are obtained by RoughEventDetection. The users for each grid point are extracted from all users by checking of intersection with each of grids and each of users. If a user intersects a grid, the user is extracted as belonging to the grid points, which are included in the grid. ConcreteEventDetection is a function to detect event concretely by using concrete grid points associated with users generated from the result of GetConcreteGridPoint. At the end of the loop, the function called ShowEventsDetected is executed to show the DetectedResult. In the ShowEventsDetected, detected result processed by the two functions of RoughEventDetection and ConcreteEventDetection. DetectedResult is composed of a set of grid points are presented to RT-SQUED system’s users.

Algorithm 2 shows the details of our two phase event detection procedures.

RoughEventDetection is a function to perform in the first rough phase. In the RoughEventDetection, DetectedRoughInfo is first initialized to store the return value. Then, for each grid and each user the intersection between the grid and the user region is iteratively checked. In this check, MinimumBoundingRectangle is created from the user triangle instead of the use of the user triangle due to rough checking. GetMinBoundingRect is a function to create a smallest rectangle, called MBR, surrounding the user triangle, which is described in the previous section. An MBR is created from the minimum and maximum values among three triangle-endpoints for each coordinate. After that, checking whether MBR intersected grid using CheckIntersect. CheckIntersect is a function to check if a MBR is intersected with a grid or not. This function uses the two coordinates of MBR and grid to check the intersection. If MBR is intersected with a grid, information on the corresponding the grid and the user is added to DetectedRoughInfo. Finally, RoughEventDetection return the DetectedRoughInfo.

ConcreteEventDetection is a function to perform the second concrete phase. In the ConcreteEventDetection, DetectedResult is first initialized to store the return value. Then, for each gridPoint and each user, it is checked whether the gridPoint is included in the user triangle, by using CheckInclusion. CheckInclusion is a function to check if the user triangle includes the gridPoint or not. This function uses the cross product vector for each of the sides of the user triangle. If a user triangle includes a gridPoint, UCounter is incremented. If the UCounter value is more than the given threshold, the corresponding gridPoints are added to DetectedResult. Finally, ConcreteEventDetection returns the DetectedResult.

IV. Evaluation & Discussions

We conducted an experiment to compare the accuracy and processing time with SQUED method and RT-SQUED. The data set that we used is one of the data set which we used in the case of the experiment of the previous paper.

TABLE I. SQUED method v.s. RT-SQUED

Method	Accuracy (F-measure)	Processing time (sec)
SQUED method	28.6%	31.67
RT-SQUED	28.6%	3.07

We set N-interval as 1 in this experiment. Table 1 shows the result of the experiment. The processing time of RT-SQUED became a one-tenth than that of SQUED method. In addition, the deterioration of the accuracy was not observed.

In our RT-SQUED, the proposed two phases, RoughEventDetection and ConcreteEventDetection, are concurrently executed. Through the experiment, we confirmed that the computational cost could be reduced and that event detection speed became faster. Moreover, the size of the detection range can be increased more. Therefore it can detect multiple events occurred simultaneously compared with SQUED method.

Furthermore, in RT-SQUED, N-interval in the first rough phase, the frequency of the RoughEventDetection execution, can be changed. This makes it possible to appropriately adjust the detection speed and detection accuracy, depending on the situation. With the ability of adjusting the frequency according to the event search range, real-time event detection can be realized.

V. Related Work

Wentao et al. proposed iSee, a detection system using the smartphone [4]. iSee uses two sensors, a GPS and a geomagnetic sensor. A user needs to swipe his/her smartphone screen toward the event location. iSee then acquires the user's location and the device direction to estimate the actual event direction using the swipe direction. Although the idea is similar to our SQUED, users need to swipe their smartphone screen. A SQUED user only needs to point his/her smartphone toward an event location. Moreover, the swipe direction contains much more noise than geomagnetic sensor data does.

Tran et al. proposed an algorithm that could recognize actions such as walking and running from videos [5]. It can detect events in a crowded video scene. In this approach, the point of intersection of the characteristic point of the object between each frame is extracted in order to recognize a movement trace. Wang et al. proposed an algorithm that could recognize other types of events, such as a parade and rock-climbing, from videos [6]. In their method, the local characteristic of the frame is replaced with a letter to detect an event more efficiently.

Tong Qin et al. proposed an crowdsourcing based event reporting system using smartphones with accurate localization and photo tamper detection [7]. In their system, it can identify the event summary and the event location by using event photographs, event descriptions and sensor data by smartphones of system users. Their system also supports the accuracy degradation due to tampering with photographs.

VI. Concludion

In this paper, we proposed a Real-Time Smart and Quick Unexpected-Event Detect (RT-SQUED) method to detect and localize unexpected events in real-time, assuming crowd sourcing with smartphone devices. We described how RT-SQUED solves SQUED's problem by using the proposed two-phase processing for the real-time event detection.

In the future, we will improve the detection method further and build a new system using RT-SQUED in order to perform the event detection in real-time.

Acknowledgment

This work was partially supported by JSPS 24300039.

References

[1] A. Harma, M. McKinney, and J. Skowronek, "Automatic surveillance of the acoustic activity in our living environment," in IEEE ICME 2005, July 2005, pp. 4 pp.6–8.

[2] A. F. Smeaton and M. McHugh, "Towards event detection in an audio-based sensor network," in ACM, ser. VSSN '05, 2005, pp. 87–94.

[3] T. Yamamoto, K. Oku, H.-H. Huang, and K. Kawagoe, "Squed: A novel crowd-sourced system for detection and localization of unexpected events from smartphone-sensor data," in IEEE/ACIS ICIS 2015, June 2015, pp. 383–386.

[4] R. W. Ouyang et al., "If you see something, swipe towards it: Crowd-sourced event localization using smartphones," in ACM, ser. UbiComp '13, 2013, pp. 23–32.

[5] D. Tran, J. Yuan, and D. Forsyth, "Video event detection: From sub-volume localization to spatiotemporal path search," Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 36, no. 2, 2014, pp. 404–416.

[6] F. Wang, Z. Sun, Y.-G. Jiang, and C.-W. Ngo, "Video event detection using motion relativity and feature selection," Multimedia, IEEE Transactions on, vol. 16, no. 5, 2014, pp. 1303–1315.

[7] T. Qin, H. Ma, D. Zhao, T. Li, and J. Chen, "Crowdsourcing based event reporting system using smartphones with accurate localization and photo tamper detection," in Big Data Computing and Communications, ser. Lecture Notes in Computer Science, 2015, vol. 9196, pp. 141–151.

Natural User Interaction Requires Good Affordance When Using a Head-Mounted Display

Takayuki Miura, Shoma Urakawa, Mitsuharu Isojima, Jianing Yu, Akihito Yoshii, Tatsuo Nakajima
Department of Computer Science and Engineering
Waseda University
Tokyo, Japan
emails: {t.miura, shoma_urakawa, mitsuharu_isojima, yu_jianing, a_yoshii, tatsuo}@dcl.cs.waseda.ac.jp

Abstract—**Natural user interface (NUI) devices have become commercially common in various types of Virtual reality (VR)-based services such as video games and public attractions. Recently, a type of head-mounted display (HMD) such as Oculus Rift also attracts the industry with the possibility of developing new types of emerging VR-based services. In this paper, we report that appropriate affordances are necessary to use respective NUI devices, particularly when a user wears an HMD, which implies that different affordances are necessary for different NUI devices. We have developed a preliminary case study to use different NUI devices, where a user wears an HMD and navigates the interaction with the case study service. We conducted an experiment to investigate the relationship between the affordances and the NUI devices to extract useful insights to develop future VR-based services that use NUI devices and HMDs. The results suggest that it is important to consider the differences of NUI devices for the affordance design to navigate VR-based services.**

Keywords-NUI; HMD; VR; Affordance; Mental model.

I. INTRODUCTION

Recently, virtual reality (VR) technologies have revived because inexpensive and practical commercial head-mounted displays (HMD) such as Oculus Rift are used [12]. Thus, various types of VR-based services [13] are easily developed and can be commercially used. For example, in Japan, some VR-based attractions using an HMD have attracted many people [14].

To offer desirable interactive users' experiences, natural user interaction (NUI) devices such as Microsoft Kinect [15] or Leap Motion [16] are widely used. Many VR-based games have been developed, and they assume to use these NUI devices because these devices offer an immersive user experience through the natural interaction, which we use in our daily life, such as hand gesture and arm movement [2], with the virtual world without prior knowledge. However, it is assumed that using the current NUI devices causes a gap between the ideal situation and reality. In particular, using NUI devices with an HMD may cause a new problem. Yang and Pan have reported that MS Kinect fails to track a user's body when the user does not have enough experience with an HMD [11]. Additionally, Sabir, Stolte, Tabor, and O'Donoghue show that poor performance when using NUI devices has been found if the users have little practice with the NUI devices [5].

As claimed in [9], considering the affordance is effective for VR-based services when used with an HMD. Thus, we believe that those types of problems occur because different NUI devices require proper affordances to use them, particularly an HMD. When using NUI devices, it is usually assumed that a user can easily find where the devices are and how to navigate them, but the devices cannot be seen when the user wears an HMD. In computing environments, various commodity NUI devices will be used to develop new VR-based services; thus, the described issues will soon become a more serious problem. We must also investigate what types of affordance are appropriate. We discuss two types of affordance: inherent and augmented affordances, which we defined based on the inherent and augmented feedforward, as proposed in [10]. Furthermore, we would like to study whether different features of NUI devices affect the appropriateness of the types of affordance. Our research question is that a different NUI device requires a different affordance. This research question is the foundation in the research area of VR, HMD, and NUI devices, but only few studies were mentioned because of the rapid development of this research area.

In this paper, we have developed a simple VR-based service to investigate the above issues as a case study and demonstrate how we can design proper affordances for respective NUI devices. The extracted insights from our experiment are useful for designing future NUI devices and VR-based services.

This paper consists of the following sections. Section II shows the background of our study and the issues that we must investigate in our study. In Section III, we explain some issues of designing affordances and how we tackle these issues. Section IV presents some related work of this study. Section V illustrates a prototype service as a case study that we developed to investigate our research question. Section VI presents our conducted experimental design, and Section VII shows the results and discussions of the experiment. Section VIII presents the conclusion and future direction of our study.

II. Background

A well-designed service requires a good mental model for navigation [3]. Traditional VR-based services typically use special and dedicated NUI devices that are developed only for the services [1]. Thus, the NUI devices usually fit well for the mental models of the services. However, in ubiquitous computing environments, we would like to adopt cheap and available NUI devices, such as MS Kinect and Leap Motion, to easily deploy the new VR-based services.

One of the potential pitfalls is that the NUI devices may cause gaps between the mental model and the assumption of the NUI devices, although it is desirable that these services can be used with any NUI device to increase the portability. For example, a user can navigate services with the movement of his/her entire body using MS Kinect, whereas he/she must assume to use only one hand to navigate the services when using Leap Motion.

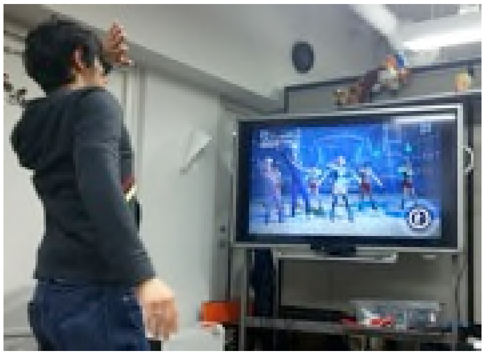


Figure 1. A Play Scene of Dance Evolution

Some concrete examples of using NUI devices are “Dance Evolution” [17] and “Nike+ Kinect Training” [18], where the video games are played on Xbox 360 using MS Kinect. “Dance Evolution” is a dance game, as shown in Figure 1, and “Nike+ Kinect Training” is an exercise game, as shown in Figure 2. In both games, the movement of each player’s body is tracked by MS Kinect, and some players can compete for scores in the games. The players easily play the games by following the visual instructions on the screen, but when we assume that MS Kinect is replaced by Leap Motion, no player may play the games well because the presented visual instructions do not afford the operations of the games with Leap Motion. This problem does not apply in these games because the games were developed to be currently operated only with MS Kinect. However, for general-purpose VR-based services that can be operated using several types of NUI devices, to overcome the issues of using various types of NUI devices for VR-based services, we require a new solution to use a VR-based service.

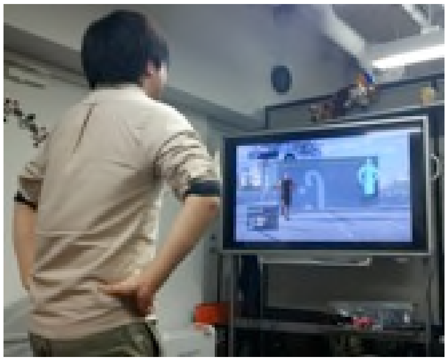


Figure 2. A Play Scene of Nike+ Kinect Training

III. Affordances and Our Research Goal

Our solution is to offer affordances to help construct the mental model to navigate the VR-based services. In this paper, we use the term “affordance” with the meaning of “perceived affordance”. A typical affordance is the knob of a door; we usually know how we can open the door without any instruction when we look at the form of the doorknob.

In other words, affordances are the functions that provide the critical clues required for their proper operation. Additionally, affordances can be used to navigate human behavior [6]. However, the following research questions remain: whether respective affordances are necessary for different NUI devices, and what types of affordances must be offered.

In this study, based on [10], we define two types of affordances: inherent affordance and augmented affordance. The inherent affordance makes us understand how we use a VR-based service based on the UI elements’ shapes, positions, etc. The definition of the affordance is widely used when designing daily objects [3]. The augmented affordance uses images or words to make us understand how we use a VR-based service. Investigating these two types of affordances enables us to extract useful insights when designing affordances for future VR-based services.

IV. Related Work

Terrenghi, Kirk, Sellen, and Izadi show that each interface creates a different affordance in [8]. In this paper, the authors asked the participants to perform a puzzle task and a task to sort photos, where each task was performed with the two following methods: using physical puzzle pieces or photos and using their digital forms, which could be operated through a touch panel, as shown in Figure 3 and 4. The result of their study is that even with identical tasks, the affordances of the respective interfaces appear differently.

In [7], Shin, Kim, and Chang studied the usability of two devices in VR-based services with HMD. They asked the participants to play a race game and an action game as shown in Figure 5 and 6, respectively, with two different types of controllers: Hydra, which must be grasped to play the game, and MS Kinect. The results of the experiment show that even in identical games, the difference in controller devices affects the impression that the users feel.

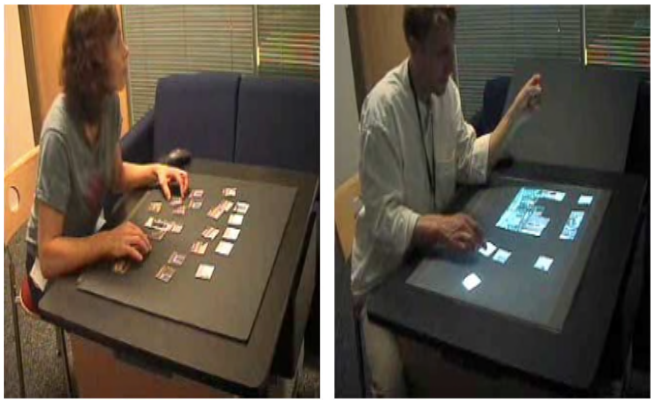


Figure 3. Puzzle Task in Two Styles [8]

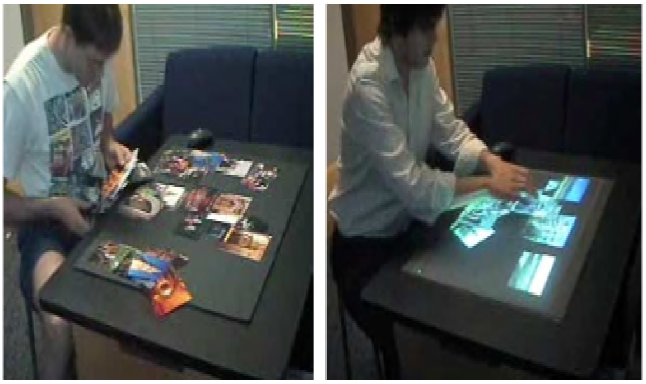


Figure 4. Sorting Photos in Two Styles [8]

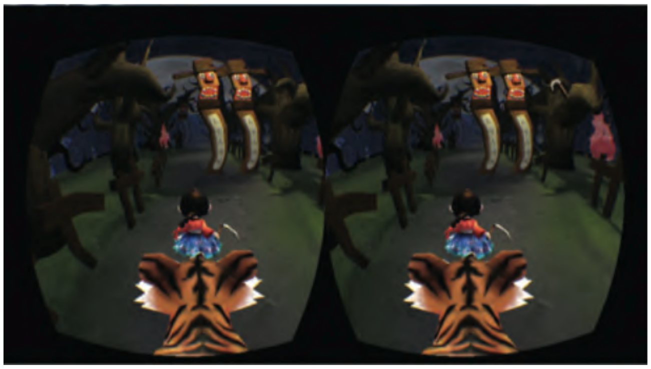


Figure 5. Race Game [7]

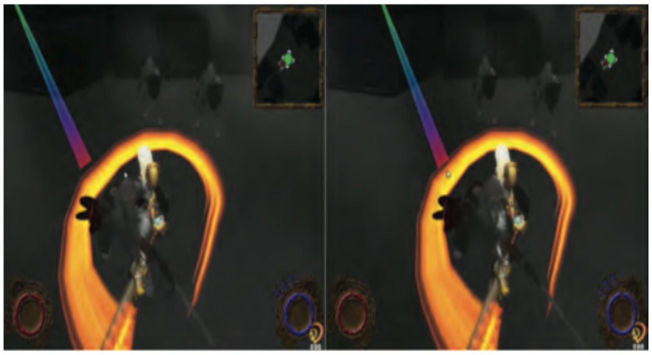


Figure 6. Action Game [7]

Thus, it is desirable to offer different affordances for each controller device to increase the usability of VR-based services when an HMD is used.

V. A Case Study

We have developed a VR-based photo viewer service as a case study, which is illustrated in Figure 7, to demonstrate the proposed ideas.

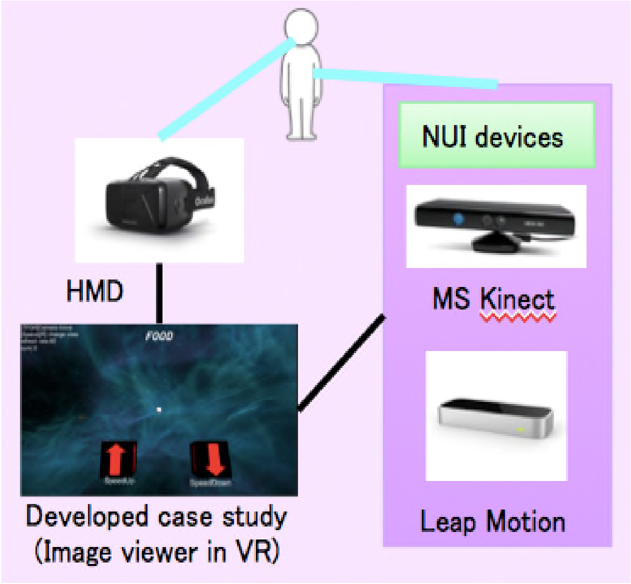
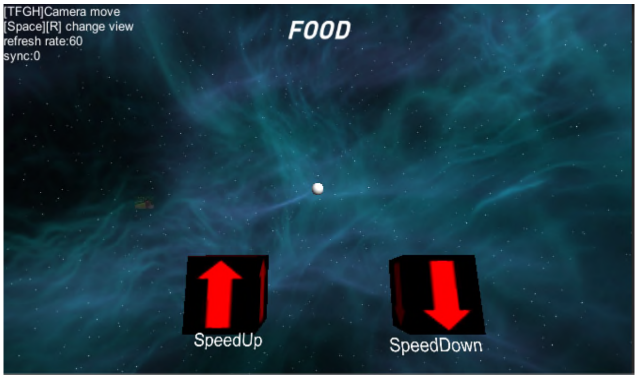


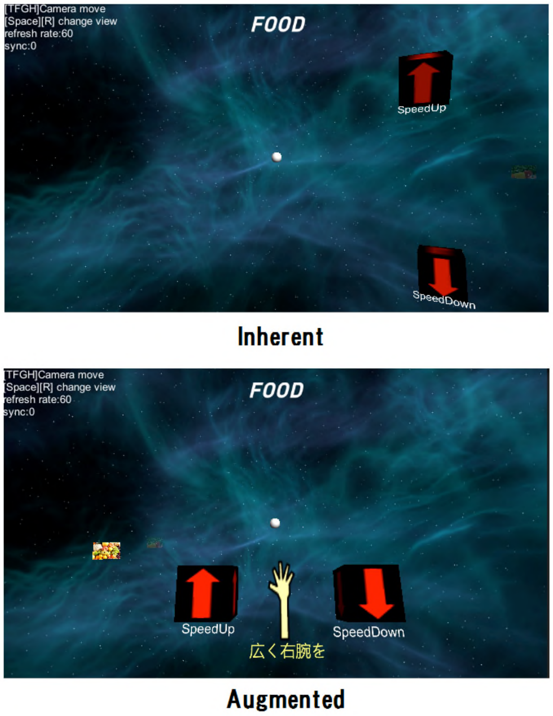
Figure 7. Construction of the image viewer we developed



[TFGH]Camera move
[Space][R] change view
refresh rate:60
sync:0

Figure 8. No Affordances for both MS Kinect and Leap Motion

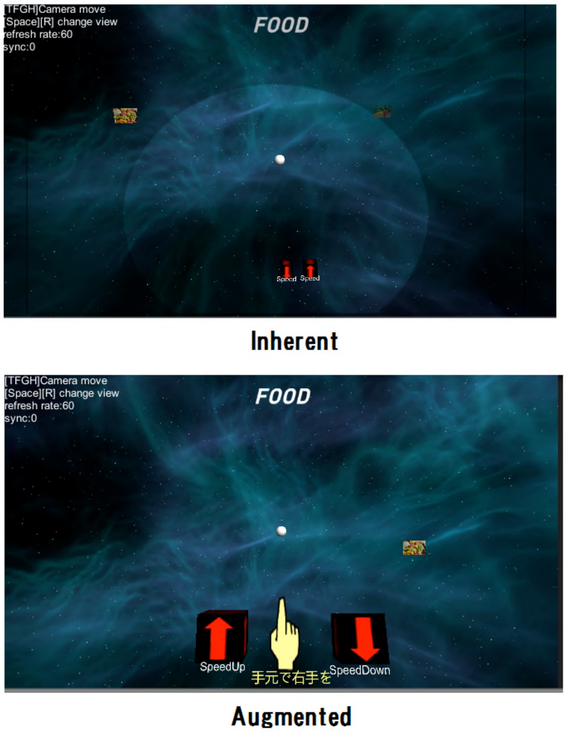
Figures 8, 9 and 10 show the screen captures of the services that are displayed in an HMD. The small white sphere shown on the screen represents a cursor that a user can navigate using his/her motion. Several photos rotate around in user's sight. At the bottom part of the screen, there are two arrow objects, which indicate "Speed up" and "Speed Down". A user adjusts the photo's moving speed by putting the moving cursor on these objects. By putting the cursor on a photo for a time period, the size of the photo is enlarged.



Inherent

Augmented

Figure 9. Inherent and Augmented Affordances for MS Kinect



Inherent

Augmented

Figure 10. Inherent and Augmented Affordances for Leap Motion

The current case study assumes to use either MS Kinect or Leap Motion as NUI devices. We designed the inherent and augmented affordances for the respective NUI devices. The service can be used without presenting these affordances.

MS Kinect tracks the positions of a user's body. In this case study, the position of a user's right hand is captured to move the cursor. Conversely, Leap Motion tracks the positions of the joints of a user's hand. In this case study, the position of the back of a user' hand is detected to move the cursor. Figure 8 is a screenshot when no affordance is shown. Figure 9 shows the screens that represent the inherent and augmented affordances for MS Kinect, and Figure 10 presents those for Leap Motion.

In the experiment of this case study, we conducted user studies for the following five combinations.

(1) *No affordance + MS Kinect or Leap Motion:* the "Speed Up" and "Speed Down" objects are shown at the bottom of the screen.

(2) *Leap Motion + Inherent affordance:* The "Speed Up" and "Speed Down" objects are represented with smaller sizes than those in the service with no affordance, and the region that a user can move his/her hand is visualized.

(3) *Leap Motion + Augmented affordance:* A picture of a hand and the sentence "Right Hand here" are displayed at the bottom of the screen.

(4) *MS Kinect + Inherent affordance:* The positions of "Speed Up" and "Speed Down" objects are shown on the top and bottom right side of the screen.

(5) *MS Kinect + Augmented affordance:* A picture of a hand and the sentence "Use Right Arm Widely" are displayed at the bottom of the screen.

VI. Experiment Design

In this study, we performed an experiment to investigate the above combinations. A participant selects two photos in each combination. The word "select" indicates enlarging the photos by putting the cursor on the photos for a period of time. In this experiment, we investigated the differences when there is an affordance or not and when two types of affordances are presented. We also investigated the situations when the participant knows what NUI device he/she uses and when he/she does not know which NUI device is used.

In this experiment, twelve participants with ages of 21-54 participated. Figure 11 shows one actual scene during the experiment.

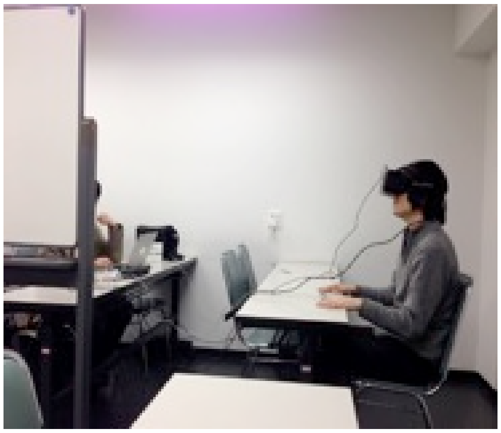


Figure 11. A Scene in Our Experiment

The experiment for one person took approximately 20-40 minutes. We conducted the semi-structured interview for them after the experiment.

VII. Design Implication

In the experiment, most participants basically showed positive attitudes for our proposed approach, and said, "*The presentation of the affordance helps me to navigate the service."* In the section, we presented some more detailed answers in the interview conducted after the experiment described in the previous section.

We asked the participants "*Why did you consider that the affordance helps you?"* Some of them answered *"How, what, and where to move my hand or arm could be understood easily by the affordance"* Also, when asking a question "*Why did you prefer the inherent affordance of MS Kinect?"* Some participants who prefer the inherent affordance for MS Kinect answered "*The inherent affordance for MS Kinect is effective because the position of the arrow objects indicates the necessary action for my right arm*"

One opinion for the merit of the inherent affordance for Leap Motion is that the operative range of the device is easy to understand. However, some participants said, "*The inherent affordance is not good for me*", and they presented its reason as follows; *"I cannot understand the meaning of the affordance, and it confused me."* When comparing the inherent affordance for MS Kinect and Leap Motion, most participants said the affordance for Leap Motion is better. In terms of the augmented affordance, when asking the participants *"Comparing MS Kinect with Leap Motion, do you think which augmented affordance was easy to understand?",* Most participants answered *"The difference between the affordances for MS Kinect and Leap Motion is small because the augmented affordances for both devices are similar*"

In the opinion about the inherent and augmented affordance, many participants said the augmented affordance is better than inherent one, because words and images are understandable in an easier way. Additionally, the inherent affordance offers the better effect when the participants know which NUI device is currently used, whereas the augmented affordance has a better effect when they do not know which NUI device is used.

Ideally, a VR-based service should offer a proper mental model regardless of the NUI devices, but in reality, it is difficult to navigate the service without knowing which NUI device is used. Some participants said that the inherent affordance was good, but the others said that it was not good. We think that the variations are caused by whether they consider that they can intuitively understand the offered affordances. We hypothesize that the inherent affordance for Leap Motion was preferred to the affordance for MS Kinect because the participants easily understood the visualization of the affordance, whereas the movable region of their hands was limited in Leap Motion case. This result may indicate that some features of NUI devices affect the difficulty of understanding particular types of affordances. Additionally, most participants feel that the differences between the augmented affordances for MS Kinect and Leap Motion are small because the images and sentences in the augmented affordance are easily understood for most participants. Thus, the variations among individuals are small.

When comparing the inherent affordance with the augmented one, many participants claimed the augmented affordance is better. We analyzed the reason of it is that the question in the interview asked only whether the affordance was easy to understand or not. The inherent affordance uses only objects' shapes or positions, so in terms of human abilities for understanding the world, words and images have significant advantages because these are useful tools that can be used for explaining the world. We consider that asking *"Did you think this affordance is suitable for the guidance?"*, does not mean not only the difficulty of understanding but also the accident in understanding, and the results may differ.

Finally, the participants that knew which NUI device was used during the experiment preferred the inherent affordances because they could construct proper mental models before they found the affordances. Thus, they preferred the inherent affordance that required a lower recognition load. Additionally, the participants that did not know which NUI device was used preferred the augmented affordance because the affordance that helped them to easily construct the mental model is desirable for them.

VIII. Conclusion and Future Direction

Recently, HMD and VR have become attractive options to develop emerging new entertainment services and attractions. In particular, video games will use them to offer more immersive game experiences. NUI devices allow the games to be naturally interacted. However, there is no sufficient discussion on how to offer affordances for different NUI devices. This paper shows that each NUI device requires a different affordance when the VR-based services are used with HMDs.

In the next step, we will investigate a more systematic design guideline for affordances based on the insights of the current experiment. For example, as shown in [4], using multiple types of affordances together may offer a better result because of different effects. We will also attempt to discuss how to use other NUI devices to expand our current insights. In addition, we should consider other VR-based services to deepen our study.

In the future, many types of NUI devices will be available to develop advanced VR-based services. However, if the developers must consider different affordances for respective NUI devices, it may become troublesome.

References

[1] T. Mazuryk and M. Gervautz, "Virtual Reality History, Applications, Technology and Future", TR-186-2-96-06, Institute of Computer Graphics and Algorithms, Vienna, University of Technology, 1996.

[2] A. Macaranas, A. Antle, and E. B. Riecke, “Three Strategies for Designing Intuitive Natural User Interfaces”, In Proceedings of Extended Abstracts of the Designing Interactive Systems (ACM DIS) Conference, 2012.

[3] D. Norman, “The Design of Everyday Things”, The MIT Press; revised and expanded, 2014.

[4] D. Norman, “Affordance and Design”, http://www.jnd.org/dn.mss/affordances_and.html, accessed 09/29/2015.

[5] K. Sabir, C. Stolte, B. Tabor, and S. I. O’Donoghue, “The Molecular Control Toolkit: Controlling 3D Molecular Graphics via Gesture and Voice”, In Proceedings of the International Symposium on Biological Data Visualization, 2013, pp. 49-56.

[6] M. Sakamoto and T. Nakajima, “In Search of the Right Abstraction for Designing Persuasive Affordance towards a Flourished Society”, In Proceedings of the 9th International Conference on Design and Semantics of Form and Movement, 2015., pp. 251-260.

[7] S. Shin, S. Kim, and J. Chang, “An Implementation of the HMD-Enabled Interface and System Usability Test”, ISSSG 2014, 2014, pp. 183-193.

[8] L. Terrenghi, D. Kirk, A. Sellen, and S. Izadi, “Affordances for Manipulation of Physical versus Digital Media on Interactive Surface”, HCI 2007 Proceedings, Novel Navigation, 2007, pp. 1157-1166.

[9] T. R. Corte, M. Marchal, G. Cirio, and A. Lecuyer “Perceiving affordances in virtual reality: influence of person and environmental properties in perception of standing on virtual grounds”, Virtual Reality vol.17, no.1, Springer-Verlag London, 2013, pp. 17-28.

[10] S. A. G. Wesveen, J. P. Djajadiningrat, and C. J. Overbeeke, “Interaction Frogger: A Design Framework to Couple Action and Function through Feedback and Feedforward”, In Proceedings of the 5th conference on Designing interactive systems, 2004, pp. 177-184.

[11] X. Yang and L. Pan, “Navigating the Virtual Environment Using Microsoft Kinect”, accessed 12/08/2015.

[12] http://www.oculus.com/, accessed 12/07/2015.

[13] http://share.oculus.com/, accessed 12/07/2015.

[14] http://weekly.ascii.jp/elem/000/000/278/278687, accessed 12/07/2015.

[15] http://www.microsoft.com/en-us/Kinectforwindows/, accessed 12/07/2015.

[16] https://www.leapmotion.com/, accessed 12/07/2015.

[17] http://www.xbox.com/ja-JP/Marketplace/SplashPages/danceevolution, accessed 12/07/2015.

[18] http://www.xbox.com/ja-JP/Marketplace/SplashPages/nike-kinect-training, accessed 12/12/2015.

Two Kinds of Audio Feature Selection Algorithms in Wavelet Domain

De Li, DaYou Jiang
Department of Computer Science
Yanbian University
Yanji, China
Email: leader1223@ybu.edu.cn
ybdxgxy13529@163.com

Jongweon Kim*
Department of Contents and Copyright
Sangmyung University
Seoul, Korea
Email:jwkim@smu.ac.kr

Abstract—**In this paper, we propose two new audio feature selection algorithms based on Discrete Wavelet Transform (DWT) domain. One of them is combined with Discrete Cosine Transform (DCT), the other one is combined with Mel Frequency Cepstral Coefficients (MFCCs). First, we introduce two different audio selection algorithms; second, we use the audio attack experiments to verify reliability of those algorithms. The tests show that the DWT-DCT algorithm has better stability under most of audio attacks, but not is robust to amplify attack for electronic music, while DWT-MFCCs algorithm is especially stable under amplify attack, but not is robust to mp3 compression attacks.**

Keywords-Audio feature selection; DWT; DCT; MFCCs

I. INTRODUCTION

In recent years, several front-ends have been proposed in the field of audio feature extraction. Some of them are based on short-term features, such as Fast Fourier Transform coefficients (FFTC) [1], DWT coefficients [2], DCT coefficients [3], MFCCs [4], real cepstral coefficients (RECC) [5], log filterbank energies [6], Perceptual linear Prediction (PLP) [7], log-energy, spectral flux, zero-crossing rate (ZCR) [8] and fundamental entropy. Others are based on the application of different temporal integration techniques over these short-term features. A. Meng [9] proposed a multivariate auto-regressive feature model which gives two different feature sets, the diagonal auto-regressive and multivariate auto-regressive features. P. Ruvolo [10] extended his previous work on Spectro-temporal box filters (STBFs) by proposing a hierarchical approach to combine features at multiple time scales. A. Suhaib [11] proposed a new method for audio feature extraction which is by using Probability Distribution Function (PDF). The PDF is a statistical method which is usually used as one of the processes in complex feature extraction methods such as Gaussian Mixture Models (GMM) and Principle Component Analysis (PCA). X. Y. Zhang [12] proposed a new time-frequency audio feature extraction scheme, in which features are decomposed from a frequency-time-scale-tensor. The tensor, derived from a weight vector and a Gabor dictionary in sparse coding, represents the frequency, time centre and scale of transient time-frequency components with different dimensions. I. Vatolkin [13] proposed an approach on how evolutionary multi-objective feature selection can be applied for a systematic maximisation of interpretability without a limitation to the usage of only interpretable features. In the experiments, 636 relevant low-level audio features and 566 high-level audio features were used. H. Muthusamy [14] proposed the particle swarm optimization based clustering (PSOC) and wrapper based particle swarm optimization (WPSO) to enhance the discreming ability of the features and to select the discriminating features respectively. In the experiments, MFCCs, linear predictive cepstral coefficients (LPCCs), PLP features, gammatone filter outputs, timbral texture features, stationary wavelet transform based timbral texture features and relative wavelet packet energy and entropy features were extracted.

In this paper, we proposed two audio selection algorithms in wavelet transform domain, which combined DCT and MFCCs. They have good stability to a series of audio attacks. Features of audio music are expressed as binary image after combined transformation.

The details of proposed algorithms will be addressed in following sections. In the next Section 2, we study MFCCs and DWT. Then Section 3 explains the proposed two different audio selection algorithms, the DWT-MFCCs and DWT-DCT. Section 4 shows the results of the audio attacks experiments with respect to two proposed algorithms to demonstrate the performance and stability of those algorithms. And finally, Section 5 gives the conclusions and future works.

II. BACKGROUND AND RELATED WORK

A. DWT

DWT represents an analog signal in the time-frequency domain with Sine and Cosine functions and the coefficients are calculated by using Mallat's pyramid algorithm [15]. The general procedure for DWT is illustrated in Figure.1. DWT decomposes a signal to approximation coefficients and detail coefficients by applying low-pass and high-pass filters respectively. The detail coefficients can be sent to another set of filters for further decomposition. The filterbank implementation of wavelets can be interpreted as computing the wavelet coefficients of a discrete set of child wavelets for a given mother wavelet ψ (t). In the case of the discrete wavelet transform, the mother wavelet is shifted and scaled by powers of two.

$$\psi_{j,k}(t) = \frac{1}{2^j}\psi(\frac{t-k2^j}{2^j}) \qquad (1)$$

Where *j* is the scale parameter and *k* is the shift parameter,

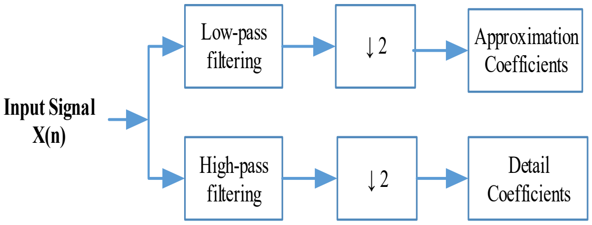


Figure 1. The general procedure for DWT

B. DCT

The DCT is a real transform that has great advantages in energy compaction. The DCT is actually shift variant, due to its cosine functions. Its definition for spectral components $f_y(k)$ is

$$f_y(k) = \begin{cases} \sqrt{\frac{1}{N}}\sum_{n=0}^{N-1} x(n)\cos\frac{k(2n+1)\pi}{2N} & k=0 \\ \sqrt{\frac{2}{N}}\sum_{n=0}^{N-1} x(n)\cos\frac{k(2n+1)\pi}{2N} & k=1,2,\ldots \end{cases} \qquad (2)$$

Where x(n) is a frame sequence of audio signal, the length of which is N. The $f_y(k)$ is the coefficients sequence obtained by applying DCT.

C. Algorithm of MFCCs feature selection

MFCCs were firstly introduced by Davis and Mermelstein in the 1980's, which is one of the most popular features for speech recognition [16]. In the past few decades, MFCC became popular parameterization method that has been developed and has been widely used in the speech technology field. MFCC analysis is similar to cepstral analysis and yet the frequency is warped in accordance with Mel-scale. The mel-cepstrum exploits auditory principles, as well as the decorrelating property of the cepstrum.

The MFCC features for a segment of music file are computed using the following procedures:

Step1: The host audio clip is loaded as signal s(n). Taking it through a preemphasis filter, after through high-pass filter, the signal s'(n) = s(n) - a × s(n-1). Set the variable a as 0.95.

Step 2: Devide the signal into short frames si(n) with frame duration as F_d and frame step as F_s. In our scheme, we set the F_d as 256 and F_s as 128, where i=1,2,⋯,Nf, and Nf is the number of frames .

Step 3: Take FFT to the signal and calculate a periodogram spectral estimate of the pow spectrum pi(k).

Step 4: Apply the mel-frequence cepstral coefficients to the pow spectral, sum the energy in each filter. In our scheme, we set the filter number as 24.

Step 5: Take the Discrete Cosine Transform (DCT) of the logarithm of all filterbank energies.

Step 6: Keep DCT coefficients 6-9, discard the rest.

Step 7: Calculate two order difference coefficients of the DCT coefficients 6-9 and calculate the mean value of parameters in each frame.

Step 8: Choose the middle 1025*2 values F_{dc} from frames number 1024 then set them number from 1 to 2050. Compute the mean value of the neighbouring two values by the following formula:

$$Fp(i) = (F_{dc}(2*i-1) + F_{dc}(2*i))/2 \quad i = 1,2,\ldots \qquad (3)$$

Step 9: Apply binarization algorithm by the following formula:

$$Fp(i) = \begin{cases} 1 & Fp(i) > Fp(i+1) \quad i = 1,2,\ldots \\ 0 & otherwise \end{cases} \qquad (4)$$

III. TWO DIFFERENT AUDIO FEATURE SELECTION ALGORITHMS

In this paper, we proposed two audio feature selection algorithms. The first one DWT-DCT is based on dual domains, while the second one is using MFCCs, which indicates the short time scale features. The feature values are finally expressed as binary image with size 32 × 32. So the audio features are represented using 1024 points.

A. DWT-DCT

The main characteristic of DWT is multi-resolution. After taking DWT to audio signal, the audio signal is decomposed into time domain and frequency domain by different scales corresponding to different frequency ranges. The approximation components indicating the low frequency components of the signal can effectively resist various attacks. The DCT has the property of decorrelation and the DC coefficient of DCT has good stability. The proposed audio feature selection algorithm based on DWT-DCT is shown in Figure 2.

The procedure of the proposed DWT-DCT feature selection algorithm is described as follows:

Step1: The host audio clip is loaded as signal s(n). Divide the signal into short frames $s_i(n)$. Each frame has 512 points, where i=1,2,⋯,Nf, and Nf is the number of frames.

Step 2: Devide each short signal $s_i(n)$ into 4 shorter frames $s_{ip}(n)$ with frame duration as F_d. where ip=1,2,3 and 4.

Step 3: Perform 2-level DWT with 1-coefficients Daubechies wavelet (Db1) to each shorter frames $s_{ip}(n)$.

Step 4: Apply 1-D DCT to approximation components, then take the DC coefficient.

Step 5: Obtain the mean value of 4 DC coefficients as the feature F_c of each frame.

Step 6: Transform F_c into integer by the following formula:

$$Fc = \begin{cases} \lfloor Fc \rfloor \text{ ,} if\ Fc - \lfloor Fc \rfloor < 0.5 \\ \lceil Fc \rceil \ otherwise \end{cases} \qquad (5)$$

Where ⌊ ⌋ and ⌈ ⌉ are the ceil function and floor function.

Step 7: Apply binarization algorithm by the following formula:

$$Fc = \begin{cases} 1 & if \ \mathrm{mod}(Fc, 2) == 0 \\ 0 & otherwise \end{cases} \qquad (6)$$

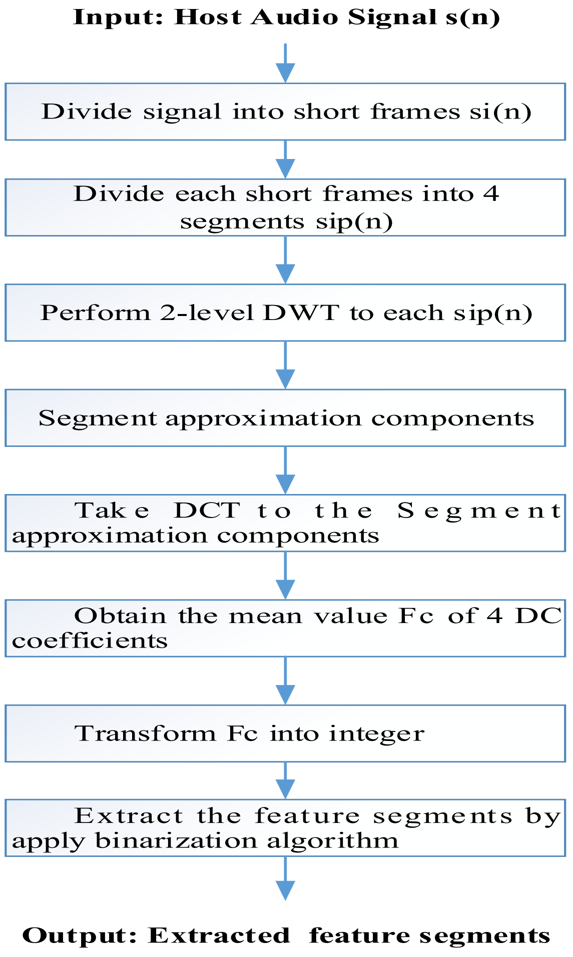

Input: Host Audio Signal s(n)
Divide signal into short frames si(n)
Divide each short frames into 4 segments sip(n)
Perform 2-level DWT to each sip(n)
Segment approximation components
Take DCT to the Segment approximation components
Obtain the mean value Fc of 4 DC coefficients
Transform Fc into integer
Extract the feature segments by apply binarization algorithm
Output: Extracted feature segments

Figure 2. Follow chart of the DWT-DCT algorithm

B. DWT-MFCCs

The design inspiration of DWT-MFCCs algorithm mainly comes from MFCCs. The low frequency components of the signal in DWT domain have higher robustness against many attacks. And the middle coefficients of MFCCs are stable. So, we can combine the DWT with MFCCs to select the feature of audio signal. The procedures of DWT-MFCCs algorithm shown in Figure 3 is virtually the same as MFCCs. The minor difference between them is that the former should perform 1-level DWT transformation on signal audio to get the approximation components for the next procedures.

IV. AUDIO ATTACK TESTS

To test our attacks,we have chosen three sound files with variant characteristics:

Electronic Music (see Figure 4).

“I am ready”: Pop music by Bryan Adams (see Figure 5).

“Mark the knife”: Jazz music by Westlife (see Figure 6).

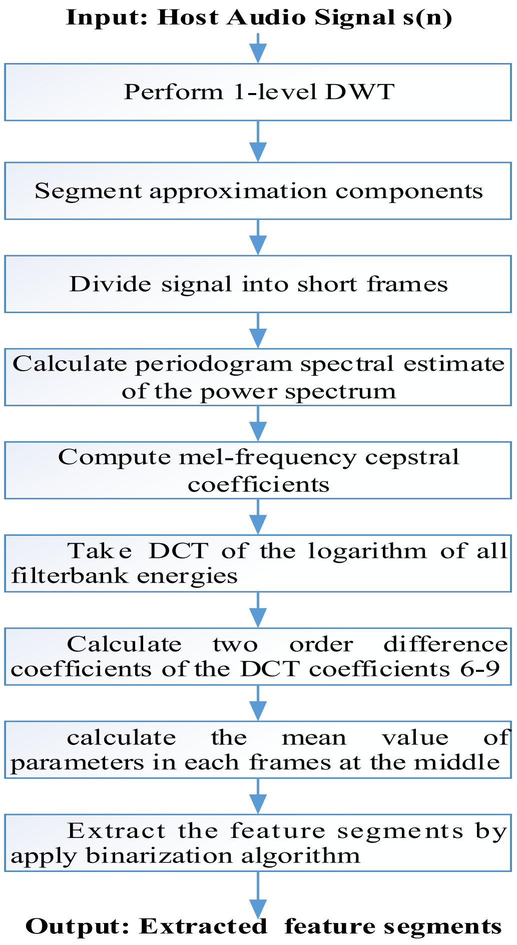

Input: Host Audio Signal s(n)
Perform 1-level DWT
Segment approximation components
Divide signal into short frames
Calculate periodogram spectral estimate of the power spectrum
Compute mel-frequency cepstral coefficients
Take DCT of the logarithm of all filterbank energies
Calculate two order difference coefficients of the DCT coefficients 6-9
calculate the mean value of parameters in each frames at the middle
Extract the feature segments by apply binarization algorithm
Output: Extracted feature segments

Figure.3. Follow chart of the DWT-MFCCs algorithm

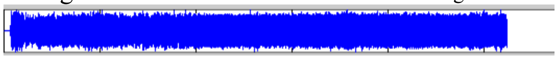


Figure 4. Electronic music

Figure 5. Pop music

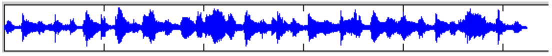


Figure 6. Jazz music

A. Attacks type

All of the audio signals in the test are music with 16bits/sample, 44100Hz sample rates, and 12 seconds. In order to illustrate the robustness of the algorithm, common signal attacks and audio stir-mark attacks are used. The following attacks are chosen:

Common signal attacks:

Noise addition: White noise with 20 dB of the power is added.

Delay: A delayed copy of the original is added to it. Set the delay time of 50 ms and a decay of 10%.

Echo: An echo signal with a delay of 50 ms and a decay of 10% is considered.

Re-quantization: Re-quantization of a 16-bit audio signal to 8-bit and back to 16-bit. Re-quantization of a 16-bit audio signal to 32-bit and back to 16-bit.

Re-sampling: the original audio signal is down sampled to 22050Hz and the up sampled back to 44100Hz.

MPEG compression: The Adobe Audition 3.0 was used to perform coding and decoding with bit rates 128Kbit and 64Kbit.

Audio Stir-mark attacks:

Noise addition: add white noise addbrumm_100.

Low-pass filtering: the resistor capacitor circuit (RC) low-pass filter with cutoff frequency of 11025 Hz is applied.

Addsinus: Add Sinus attack with frequency in 900 Hz and amplitude as the value of 1300.

Amplify: set the volume down to 50%.

In the study, reliability was measured as the bit error rate (BER) of the extracted feature, and its definition is:

$$BER = \frac{BE}{NL} \times 100\% \tag{6}$$

Where BE and NL are respectively the number of erroneously detected bits and the gross feature bits.

B. *Experimental results*

Tables 1 and 2 show the results under the audio attacks mentioned above.

TABLE I. ATTACKS RESULTS FOR DWT-DCT ALGORITHM (BER)

attacks	***Pop***	***Electronic***	***Jazz***
addnoise (20dB)	0	0.0127	0.001
addbrumm_100	0.001	0.0059	0.0029
addsinus	0	0.0166	0.002
amplify (0.5)	0.0068	0.1172	0.0301
echo_50_10%	0.001	0.0176	0.0059
delay_50_10%	0	0.042	0.0273
rc_lowpass	0	0.001	0.001
Requantization 8	0	0.0029	0
Requantization 32	0	0	0
resampling	0	0	0
mp3 128kbps	0.0088	0.2041	0.0447
mp3 64kbps	0.0088	0.2041	0.0457

As shown in Table 1, since that the DC values have nice stability, the DWT-DCT algorithm resists most attacks, especially under add noise, add sinus, requantization and re-sampling attack. After mp3 compression, the feature values didn't change a lot for Pop music and Jazz music. Only for Electronic music, the robustness under amplify attack and mp3 compression attacks is lower than the other music.

TABLE II. ATTACKS RESULTS FOR DWT-MFCCS ALGORITHM (BER)

attacks	***Pop***	***Electronic***	***Jazz***
addnoise (20dB)	0.0264	0.0417	0.0296
addbrumm_100	0.0176	0.0078	0.0269
addsinus	0.008	0.032	0.0164
amplify (0.5)	0	0	0
echo_50_10%	0.0352	0.0521	0.0449
delay_50_10%	0.0469	0.0703	0.0623
rc_lowpass	0	0	0.002
Requantization 8	0.0137	0.0234	0.0273
Requantization 32	0	0	0
resampling	0	0	0
mp3 128kbps	0.3554	0.5344	0.3961
mp3 64kbps	0.3554	0.5344	0.3971

As shown in Table 2, the DWT-MFCCs algorithm is robust to most attacks, especially under amplify, rc-low-pass filtering and re-sampling attack. The limitation is that it cannot resist to MP3 compression. Because compression and reductive process, though the format of audio changed back to .wav, the length of the signal has been changed.

Comparing the experimental results of DWT-MFCCs algorithm with the DWT-DCT algorithm, we can find that the DWT-DCT algorithm is generally much better under these attacks. But for audio segment like the electronic music tested in the paper, the audio feature selection algorithm using the DWT-MFCCs algorithm can be more robust under amplify attack.

V. CONCLUSIONS

We proposed two audio feature selection algorithms based on wavelet domain. To verify the reliability of the algorithms, three different audio music signals were tested under a series of attacks including common signal attacks and audio stir-mark attacks. The experimental results show that the DWT-DCT algorithm has better stability than DWT-MFCCs to most attacks, and the DWT-MFCCs is more robust under amplifying attack.

The proposed algorithms have desynchronization problems and the audio features will have a big change after the desynchronization attacks. The future work will be focused on solving the problem of synchronization.

ACKNOWLEDGEMENTS

This research project was supported by Ministry of Culture, Sports and Tourism (MCST) and from Korea Copyright Commission in 2015.

REFERENCES

[1] D. Megias, J. Serra-Ruiz, and M. Fallahpour. “Efficient self-synchronized blind audio watermarking system based on time domain and FFT amplitude modification”.Signal processing vol.90, 2010, pp.3078-3092.

[2] S. G. Mallat. “A wavelet tour of signal processing”. Academic, New York. 1999

[3] H. T, Hu and L. Y. Hsu. “Robust, transparent and high-capacity audio watermarking in DCT domain”. Signal Processing vol.109, 2015, pp.226-235.

[4] L. C. Jimmy and G. A. Ascensiόn. “Feature extraction based on the high-pass filtering of audio signals for Acoustic Event Classification”. Computer Speech and Language vol.30, 2015, pp.32-42.

[5] B. Gold and N. Morgan. “Speech and audio signal processing: Processing and Perception of Speech and Music.Wiley”.2000

[6] X. Zhuang, et al. “Real-world acoustic event detection”. Pattern Recognition Letters. vol.31, 2010, pp.1543-1551.

[7] J. Portel, et al. “Non speech audio event detection”. In: IEEE Int. Conf. On Acoustics, Speech, and Signal Processing (ICASSP), 1973-1976.

[8] A. Temko and C. Nadeu. “Classification of acoustic events using SVM-based clustering schemes”. Pattern Recognition vol.39, 2006, pp.684-694.

[9] A . Meng, P. Ahrendt and J. Larsen. “Temporal feature integration for music genre classification”. IEEE Trans. Audio Speech Language Process. vol.15, 2007, pp.1654-1664.

[10] P. Ruvolo, I. Fasel, and J. R. Movellan. “A learning approach to hierarchical feature selection and aggregation for audio classification”. Pattern Recognition Letters. vol.31, 2010, pp.1535-1542.

[11] A. Suhaib, et al. “Audio feature extraction using probability distribution function”. AIP Conf. Proc. Penang, Malaysia, May 28-30, 2014.

[12] X. Y. Zhang, et al. “Time–frequency audio feature extraction based on tensor representation of sparse coding”. Electronic Letters. Vol.51, no.2, 2015, pp.131-132.

[13] I. Vatolkin, et al. “Interpretability of Music Classification as a Criterion for Evolutionary Multi-objective Feature Selection”.4th International Conference, EvoMUSART 2015, Copenhagen, Denmark, April 8-10, 2015, Proceedings, pp.236-243.

[14] H. Muthusamy, et al. “Particle Swarm Optimization Based Feature Enhancement and Feature Selection for Improved Emotion Recognition in Speech and Glottal Signals”. PLoS One. vol.10, no.3, 2015, pp.1-20.

[15] S. Mallat. “A theory for multi-resolution signal decomposition: the wavelet representation”. IEEE Trans. Pattern Anal.vol.11, no.7, 1989, pp.674-693.

[16] P. Beyerlein, et al. “Large vocabulary cretinous speech recognition of broadcast news- the Philip/RWTH approach”, speech Commun.2002.

A Fast Mode Determination Algorithm Using Spatiotemporal and Depth Information for High Efficiency Video Coding (HEVC)

Kyung-Soon Jang, Jong-Hyeok Lee
Department Of Computer Engineering
SunMoon University
A-san, Rep. of Korea,
{ks.jang, jh.Lee}@mpcl.sunmoon.ac.kr

Chan-seob Park, Byung-Gyu Kim
Department Of Computer Engineering
SunMoon University
A-san, Rep. of Korea,
{cs.Park, bg.kim}@mpcl.sunmoon.ac.kr

Abstract— **High Efficiency Video Coding (HEVC) is a successor to the H.264/AVC standard and is the newest video coding standard using a quad-tree structure with three block types of a coding unit (CU), a prediction unit (PU), and a transform unit (TU). HEVC uses all possible depth levels to determine the lowest RD-cost block. Thus, HEVC is more computationally complex than the previous H.264/AVC standard. To overcome this problem, an early skip and merge mode detection algorithm is proposed using spatiotemporal and depth information. Experimental results show that the proposed algorithm can achieve an approximate 40% time saving with a random-access profile while maintaining comparable rate-distortion performance, compared with HM 12.0 reference software.**

Keywords-High Efficiency Video Coding (HEVC); Early Skip Mode Detection; Inter Prediction; Depth; Coding Unit (CU).

I. INTRODUCTION

High Efficient Video Coding (HEVC) is the latest international video coding standard issued by the Joint Collaborative Team on Video Coding (JCT-VC) [1], which is a partnership between the ITU-T Video Coding Experts Group (VCEG) and the ISO/IEC Moving Picture Experts Group (MPEG), two prominent international organizations that specify video coding standards [2].

Increasing demands for high quality full high definition (Full HD), ultra high definition (UHD), and higher resolution video necessitate bitrate savings for broadcasting and video streaming. HEVC aims to achieve a 50% bitrate reduction, compared with the previous H.264/AVC standard, while maintaining quality.

HEVC is based on a coding unit (CU), a prediction unit (PU), and a transform unit (TU). The CU is a basic coding unit analogous to the concept of macroblocks in H.264/AVC. However, a coding tree unit (CTU) is the largest CU size that can be partitioned into 4 sub-CUs of sizes from 64x64 to 8x8. Figure 1 shows an example of the CU partitioning structure. This flexible quad-tree structure leads to improved rate-distortion performance and also HEVC features for advanced content adaptability. The PU is the basic unit of inter/intra-prediction containing large blocks composed of smaller blocks of different symmetric shapes, such as square, and rectangular, and asymmetric. The TU is the basic unit of transformation defined independently from the PU, but whose size is limited to the CU, to which the TU belongs. Separation of the block structure into three different concepts allows optimization according to role, which results in improved coding efficiency [3], [4]. However, these advanced tools cause an extremely high computational complexity. Therefore, a decrease in the computational complexity is desired.

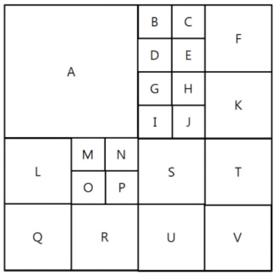


Figure 1. An example of the CU partitioning structure.

Previous work has focused on reduction of the computational complexity. Pai-Tse *et al.* [6] proposed a fast zero block detection algorithm based on SAD values using inter-prediction results. Features of the proposed algorithm were applied to different HEVC transform sizes. A 48% time saving for quantization parameter (QP) = 32 was achieved. Zhaoqing Pan *et al.* [7] proposed a fast CTU depth decision algorithm using the best quad-tree depth of spatial and temporal neighboring CTUs, relative to the current CTU, for an early quad-tree depth 0 decision. Correlations between the PU mode and the best CTU depth selection were also used for a depth 3 skipped decision. A 38% time reduction for all QPs was achieved under common testing conditions. Hassan Kibeya *et al.* [8] proposed a fast CU decision algorithm for the block structure encoding process. Based on early detection of zero quantized blocks, the number of CU partitions to be searched was reduced. Therefore, a significant reduction of encoder complexity was achieved and the proposed algorithm had almost no loss of bitrate or peak signal to noise ratio (PSNR), compared with HM 10.0 reference software.

In this study, an early skip and merge mode detection algorithm is proposed using neighboring block and depth

TABLE I. CONDITIONAL PROBABILITY BETWEEN CURRENT CTU AND NEIGHBORING CTUS.

		Spatial CTUs (%)				**Temporal CTUs (%)**								
		P(O|A)	***P(O|B)***	***P(O|C)***	***P(O|D)***	***P(O|E)***	***P(O|F)***	***P(O|G)***	***P(O|H)***	***P(O|I)***	***P(O|J)***	***P(O|K)***	***P(O|L)***	***P(O|M)***
Class B	***SKIP***	70.4	69.1	70.8	69.0	53.7	52.4	51.8	52.8	51.2	51.9	52.4	52.6	51.5
	CBF	86.6	85.8	86.8	84.8	69.1	68.2	68.1	68.8	68.0	68.2	67.8	68.4	67.6
	MERGE	89.1	83.2	83.8	82.7	75.8	75.0	73.5	74.5	72.7	72.3	72.8	73.8	66.7
Class C	***SKIP***	64.3	62.7	64.6	56.0	48.2	46.7	48.7	45.7	43.0	42.9	41.9	43.9	44.5
	CBF	85.1	85.6	86.7	84.0	69.7	60.8	69.5	67.1	61.0	56.7	55.9	58.2	64.1
	MERGE	77.2	75.8	73.8	74.44	70.0	67.7	69.2	65.8	66.0	67.4	67.9	67.3	67.2

information. A statistical analysis of the proposed method is presented in Section II. Experimental results are shown in Section III. A conclusion is presented in Section IV

II. PROPOSED ALGORITHM

HEVC includes too many stages of quad tree based structure to determine the best mode. The CU range starts from 64x64 to 8x8, so HEVC has four depth levels. In each depth level, an exhaustive RD-cost analysis is performed to determine the best mode among many modes, including SKIP, INTER_2Nx2N, INTER_Nx2N, INTER_2NxN, INTER_NxN, INTER_2NxnU, INTER_2NxnD, INTER_nLx2N, INTER_nRx2N, INTRA_2Nx2N, INTRA_NxN, and PCM. Equation (1) represents a way of determining the best mode. λ_{MODE} indicates a Lagrangian multiplier in (1), and SSE is a cost function, as defined in (2).

$$J_{MODE} = SSE + \lambda_{MODE} \quad B_{MODE} \tag{1}$$

$$SSE = \sum_{i,j} (BlockA(i,j) \quad BlockB(i,j))^2 \tag{2}$$

However, if a close correlation between the current CU and other CUs can be defined, there is no need to calculate all modes. In this study, the skip flag, merge flag, and coded block flag (CBF) of neighboring CTUs are used to reduce the computational complexity.

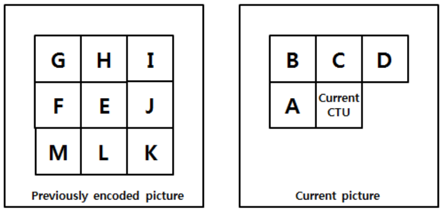


Figure 2. 13 neighboring CTUs around the current CTU.

A. Statistical Analysis

Statistical analysis was performed for identification of correlations between the current CTU and neighboring CTUs for skip and merge modes. There are thirteen neighboring CTUs around a current CTU (Figure 2). Four spatially neighboring CTUs (A ~ D) are represented as left, above-left, above, and above-right. In a previously encoded picture, there are nine temporally adjacent CTUs (E ~ M), which indicate collocated-current, collocated-left, above-left, collocated-above, collocated-above-right, collocated-right, collocated-below-right, collocated-below and collocated-below-left, respectively. Conditional probabilities include all adjacent CTUs (A ~ M) and skip and merge flag = true and

CBF = 0 are also checked. At the same time, probabilities that the current CTU, expressed alphabetically as $\mathbf{O} \in (O)$ were skip and merge flag true and CBF 0 were also checked. HM 12.0, the reference software for the HEVC standard was used. The first 10 frames of the Kimono (1920x1080), ParkScene (1920x1080), Cactus (1920x1080), BQTerrace (1920x1080), BasketballDrive (1920x1080), RaceHorses (832x480), BQMall (832x480), PartyScene (832x480) and BasketballDrill (832x480) sequences were used with QP=22, 27, 32, and 37 with a random access profile.

Statistical results showed that P(O|A ~ D), indicating a high percentage of spatial CTUs. The probabilities of each CTU having the same skip flag true were 69.1 ~ 70.8% in class B and 55.99 ~ 64.6% in class C, and also, 84.75 ~ 85.8% in class B and 84.0 ~ 86.7% in class C (Table I). Also, the probabilities of CBF 0 that CTU should be chosen as skip mode were 84.8 ~ 86.8% in class B and 84.0 ~ 86.7% in class C. In addition, the probabilities of each CTU having the same merge flag true were 82.7 ~ 89.1% in class B and 73.8 ~ 77.2% in class C. In temporal CTUs with P(O|E ~ M), percentages of each mode were 51.5% ~ 53.7% in class B and 41.9 ~ 48.7% in class B for skip flag. Also, 67.6 ~ 69.1% in class B and 56.7 ~ 69.7% in class C were recoded for CBF. Lastly, 66.7 ~ 75.8% in class B and 65.8 ~ 70.0 % in class C were recoded for merge mode.

B. Early skip and merge mode decision

If only the information mentioned above is used, loss is excessive. However, combining adjacent CTUs can reduce the loss. CTUs were divided into groups of built-in CTUs (BIC) and user defined CTUs (UDC) based on conditional probabilities, and statistical analysis was performed. Two groups were used because BICs have higher probability than UDCs, based on statistical results. Thus,

$$D = \sum_{BIC \in i} CU_i + (\sum_{UDC \in j} CU_j \quad \omega) \geq \alpha \tag{3}$$

where BIC are the left, above-left, above, and above-right CTUs in the current picture and collocated-current CTUs in the previously encoded picture. Also, UDC contains collocated-left, collocated-above-left, collocated-above, collocated-above-right, collocated-right, collocated-below-right, collocated-below, and collocated-below-left CTUs in the previously encoded picture. When CU_i and CU_j have a depth = 0, and skip flag = true or merge flag = true or CBF = 0, they are set to 1. ω is a weighting factor for UDC set to 0.75 when CU_j has the lowest RD-cost. Otherwise, ω is set to 0. α is a threshold value to specify the boundary of skip or merge modes. Lastly, Equation (3) is used for three types of calculations of skip flag (D_{SKIP}), CBF (D_{CBF}), and merge flag (D_{MERGE}). To obtain the value of D_{SKIP}, skip flag is used. For D_{CBF}, CBF is used, and for D_{MERGE}, merge flag is used.

TABLE II. STATISTICAL RESULTS ACCORDING TO α

	α	4	4.75	5	5.5	5.75	6.5
Class B	***BDBR (%)***	3.22	2.87	1.18	1.92	1.02	1.12
	ΔBitrate(%)	0.08	-0.10	-0.38	-0.13	-0.40	-0.36
	ΔPSNR(%)	-0.23	-0.16	-0.12	-0.15	-0.12	-0.12
	TS (%)	43.28	39.31	44.78	39.44	42.39	39.48
Class C	***BDBR (dB)***	3.88	3.17	1.33	2.82	1.44	1.27
	ΔBitrate(%)	1.32	0.80	0.10	0.75	0.098	0.07
	ΔPSNR(%)	-0.35	-0.29	-0.17	-0.29	-0.17	-0.16
	TS (%)	31.20	30.45	28.37	30.45	29.82	28.10

Experiments were performed to determine an optimal threshold value of α with the lowest and the second lowest RD-cost values used to obtain the optimal α. Results are shown in Table II. When α = 4, four BICs are used. Bjøntegaard difference bitrate (BDBR) values were 3.22%, and 3.88%, and TS values were 43.28% and 31.2% for classes B and C. When α = 4.75, four BICs were used and one UDC was used, BDBR values were 2.87% and 3.17% and TS values were 39.31% and 30.45% in class B and class C, respectively. When α = 5, all BICs were used. BDBR values were 1.18% and 1.44%, and TS values were 44.78% and 28.82% in classes B and C, respectively. When α 5.5, four BICs were used and two UDCs were used. BDBR values were 1.18% and 1.33% and TS values were 44.78% and 28.37%, respectively. Also, when α = 5.75, all BICs and one UDC were used. BDBR values were 1.02% and 1.44% and TS values were 42.39% and 28.32% in classes B and C, respectively. When α = 6.5, all BICs and two UDCs were used. BDBR values were 1.12% and 1.27%, and TS values were 39.48% and 28.1% in classes B and C, respectively. Good efficiency was observed when α = 5.75 in both BDBR values and TS values based on the above simulation. Therefore, 5.75 was used as a threshold value.

C. Overall Algorithm

The proposed algorithm is summarized in a flowchart in Figure 3.

Step 1. Start encoding a CTU.

Step 2. Check current depth. If depth level 0, go to Step 3. Otherwise, go to Step 8.

Step 3. Select one CTU with the lowest RD-cost value from UDC (F ~ M CTUs).

Step 4. Get all skip flags, CBFs, and merge flags from adjacent CTUs, which include all BIC (A ~ E CTUs) and one UDC selected through Step 3.

Step 5. Calculate Equation (3). If D_{SKIP} is equal to or greater than α, the current CTU is selected as skip mode. Also, go to Step 9. Otherwise, go to Step 6.

Step 6. Calculate Equation (3). If D_{CBF} is equal to or greater than α, the current CTU is selected as skip mode. Also, go to Step 9. Otherwise, go to Step 7.

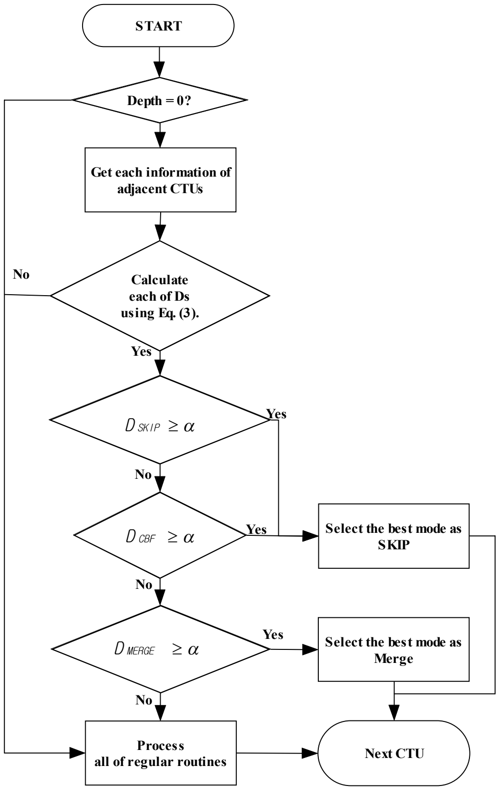

START
Depth = 0?
Get each information of adjacent CTUs
No
Calculate each of Ds using Eq. (3).
Yes
$D_{SKIP} \geq \alpha$
Yes
No
$D_{CBF} \geq \alpha$
Yes
Select the best mode as SKIP
No
$D_{MERGE} \geq \alpha$
Yes
Select the best mode as Merge
No
Process all of regular routines
Next CTU

Figure 3. Flow chart of the proposed algorithm.

Step 7. Calculate Equation (3). If D_{MERGE} is equal to or greater than α, the current CTU is selected as merge mode. Also, go to Step 9. Otherwise, go to Step 8.

Step 8. Process all regular routines.

ISBN: 978-1-61208-452-7

Step 9. Encode the next CTU.

III. EXPERIMENTAL RESULTS

The proposed algorithm was implemented on HEVC test model HM 12.0 and tested based on test conditions, configurations, and sequences recommended by JCT-VC [5]. These conditions and configurations are summarized in Table III. Performance evaluation was based on BDBR [8] and a computational complexity reduction in time saving (TS) as:

$$TS = \frac{Time(origin) \quad Time(prop)}{Time(origin)} \times 100 \qquad (4)$$

where $Time_{origin}$ and $Time_{prop}$ are the encoding times of reference software HM 12.0 and the proposed algorithm, respectively. For BDBR and TS, positive values indicated an increase and negative values a decrease.

TABLE III. TEST CONDITOINS.

CPU	Intel i5-3470 CPU @ 3.20GHz
RAM	4.00GB
OS	Microsoft Windows 7
Profile	RA-Main, LD-Main
Motion Search	TZ search
Search Range	64
Max CU Size	64x64
Max CU Depth	4
QP	22, 27, 32, 37
FEN, FDM	On

TABLE IV. SUMMARY OF ENCODING RESULTS IN RA.

Class (resolution)	Sequence	Frame count	BDBR (%)	TS (%)
A – 4K (2560x1600)	Traffic	150	1.6	45.2
	PeopleOnStreet	150	1.4	17.7
B – 1080p (1920x1080)	ParkScene	240	1.4	47.0
	Cactus	500	1.4	42.1
	BQTerrace	600	1.3	46.5
	BasketballDrive	500	1.0	36.2
C - WVGA (832x480)	BQMall	600	3.1	42.4
	PartyScene	500	1.1	30.5
	BasketballDrill	500	0.9	34.7
D - WQVGA (416x240)	BQSquare	600	0.5	36.0
	BlowingBubbles	500	1.2	29.1
	BasketballPass	500	0.5	21.2

TABLE V. SUMMARY OF ENCODING RESULTS IN LD.

Class (resolution)	Sequence	Frame count	BDBR (%)	TS (%)
B – 1080p (1920x1080)	ParkScene	240	1.0	29.7
	Cactus	500	0.5	29.2
	BQTerrace	600	0.8	36.1
	BasketballDrive	500	1.0	26.2
C - WVGA (832x480)	BQMall	600	2.5	34.7
	PartyScene	500	0.2	18.8
	BasketballDrill	500	0.9	26.3
D - WQVGA (416x240)	BQSquare	600	-0.1	18.4
	BlowingBubbles	500	0.2	16.3
	BasketballPass	500	0.3	16.1

Tables IV and V show performance results for random-access and low-delay, respectively. On average, the BDBR value was 1.3% and TS value was 31.45% in the random-access profile for A class sequences. In B class sequences, a 1.275% BDBR value was observed with a speed-up of 42.95% in random-access. In low-delay profiles, 0.825% for BDBR and 30.3% for TS were achieved.

Also, in C class sequences, BDBR values were 1.7% and 1.2% while obtaining 35.87% and 26.6% for TS with random-access and low-delay profiles respectively. In D class sequences, BDBR values of 0.73% and 0.13% with TS values of 28.77% and 16.93% with random access and low delay profiles respectively, were achieved. The proposed algorithm reduced the time required with minimal quality degradation, compared with the original encoder.

IV. CONCLUSION

An early skip and merge mode decision algorithm has been proposed based on spatially and temporally adjacent CTUs. If skip or merge mode is identified, further processes are omitted. Experimental results show that the proposed algorithm achieved an average time reduction of 34.76% in random access profile and 24.61% in low delay profile, while maintaining a comparable RD performance. The proposed method can be useful for supporting a real-time HEVC encoder implementation.

REFERENCES

[1] B. Bross, W. -J. Han, G. J. Sullivan, J. -R. Ohm, and T. Wiegand, “High efficiency video coding (HEVC) text specification draft 8,” ITU-T/ISO/IEC Joint Collaborative Team on Video Coding (JCT-VC) document JCTVC-J1003, July 2012.

[2] G. J. Sullivan, J. –R. Ohm, W.-J. Han and T. Wiegand, “Overview of the high efficiency video coding (HEVC) standard”, IEEE Trans. Circuits Syst. Video Technol., Vol. 22, no. 12, Dec. 2012, pp. 1649-1706.

[3] I. -K. Kim, J. Min, T. Lee, W. -J. Han and J. -H. Park, “Block partitioning Structures in HEVC”, IEEE Trans. Circuits Syst. Video Technol., vol. 22, no. 12, Dec. 2012, pp. 1697-1706.

[4] “Common test conditions and software reference configurations”, ITU-T/ISO/IEC Joint Collaborative Team on Video Coding (JCT-VC) document JCTVC-I1100, May 2012.

[5] P. –T. Chiang, T. S. Chang, “Fast zero block detection and early CU termination for HEVC video coding”, IEEE International Symposium on Circuits and Systems (ISCAS), May 2013, pp. 1640-1643.

[6] Zhaoping Pan, Sam Kwong, Yun Zhang, Jianjun Lei, “Fast coding tree unit depth decision for high efficiency video coding”, IEEE International Conference on Image Processing (ICIP), Oct, 2014, pp. 3214-3218.

[7] Kibeya, H., Belghith, F., Ben Ayed, M. A, Masmoudi, N., “A fast CU partitionning algorithm based on early detection of zero block quantified transform coefficients for HEVC standard”, IEEE International Conference on Image Processing, Applications and Systems (IPAS), Nov 2014, pp.1-5.

[8] G. Bjøntegaard, “Calculation of Average PSNR Differences between RD Curves”, document VCEG-M33, Austin, TX, USA, Apr. 2001.

Low Complexity Multiple Candidate Motion Estimation Using Gray Coded Bit-plane Matching

Changryoul Choi and Jechang Jeong
Dept. of Electronic and Communication Engineering Hanyang University
Seoul, Korea
e-mail: denebchoi@gmail.com & jjeong@hanyang.ac.kr

***Abstract*—In this paper, we propose low complexity motion estimation algorithms based on the Gray-coded bit-plane matching. By exploiting almost identical operations among similar but different matching error criteria, we can efficiently determine the respective candidate motion vectors. In addition, adopting multiple candidate motion estimation strategies into those candidate motion vectors and local searches around the best candidate motion vector dramatically enhance the motion estimation accuracy. Experiments were carried out for comparing the performances of the proposed algorithms with other bit-plane matching based motion estimation algorithms and full search block matching algorithm with the sum of absolute differences as well. Surprisingly, the peak signal to noise ratio difference between one of the proposed algorithms and the full search block matching algorithm is within 0.05dB on average.**

Keywords-motion estimation; bit-plane matching; video coding

I. Introduction

Due to the rapid growth of the multimedia service, the video compression has become essential for reducing the required bandwidth for transmission and storage in many applications. In video compression, the motion estimation (ME) and the motion compensation (MC) is the most crucial part since it can reduce the total video data efficiently by exploiting the temporal correlation between neighboring frames. The block matching algorithm (BMA) is adopted in many video compression standards because of its simplicity and effectiveness [1][2]. In BMA, a current frame is partitioned into small square (possibly rectangular) blocks and a motion vector is estimated within its search range in the reference frame by searching the most similar block according to some matching criterion such as the sum of absolute differences (SAD). Although the full search BMA (FSBMA) can find an optimal motion vector according to some matching criterion, its computational complexity is so huge that it is not adequate for real time applications. Therefore, many techniques including the fast searching and the fast matching algorithms have been proposed to reduce the high computational complexity of the FSBMA in the literature. Among them, there are some techniques that use different matching criteria instead of the classical SAD to make faster the computation of the matching criterion itself exploiting the bit-wise operations [3]-[12]. These algorithms are called bit-plane matching (BPM) based ME. The advantages of these techniques over the matching algorithms using the classical SAD are two-fold: fast computation of the matching criterion and reduced memory bandwidth in the interim of ME process. These techniques include one-bit transform (1BT) [3], multiplication-free 1BT [4], constrained 1BT (C1BT) [5], two-bit transform (2BT) [6], weighted 2BT [9], truncated Gray-coded BPM (TGCBPM) [7][12], weightless TGCBPM (WTGCBPM) [8], etc. Among the above algorithms that use bit-wise operations based matching criterion, TGCBPM and WTGCBPM show the best results in terms of the ME accuracy. In addition, the transforming the frame into bit-planes is relatively simple.

Although the BPM based ME succeeded in reducing the computational complexity, its ME accuracy is relatively poor compared with the SAD based ME, resulting in degraded reconstructed images. To remedy this situation, multiple candidate motion searches were proposed in [11]. In general, performing many motion searches using different matching criteria simultaneously helps improving the overall performance, but their computational complexities would increase heavily with the number of matching criteria used. However, they exploited correlation between two different matching criteria and did the dual motion searches with negligible computational complexity increase.

In this paper, we propose low complexity ME algorithms based on the Gray-coded BPM. By exploiting the similar operations between different matching error criteria, we can efficiently determine the respective candidate motion vectors. Adopting the multiple candidate motion search strategy into those candidate motion vectors and local searches around the best candidate motion vector dramatically enhance the motion estimation accuracy.

The rest of this paper is organized as follows. In Section 2, we review some previous works related to the proposed algorithm. In Section 3, we explain our proposed algorithm. Experimental results and analyses are presented at Section 4. Finally, Section 5 provides our conclusions.

II. Previous Algorithms

In this section, we review some previous works on BPM based ME and multiple candidate motion search.

A. Bit-plane Matching Based Motion Estimation

In [12], Gray-coded BPM based ME was first proposed. And, TGCBPM and its variation WTGCBPM were proposed in [7][8]. Both TGCBPM and WTGCBPM use Gray-code mapping as transforming image frames into bit-planes, which is very simple compared to other bit-plane transformation algorithms using complex filtering operations (e.g., 1BT, 2BT, C1BT, etc.). Let the gray-level of the pixel at location (*m*, *n*) be represented as follows:

$$f(m,n) = a_{K-1}2^{K-1} + a_{K-2}2^{K-2} + \cdots + a_1 2^1 + a_0 2^0 \qquad (1)$$

where a^k ($0 \le k \le K$-1) takes on either 0 or 1. Then the corresponding Gray-code representation is given by

$$\begin{aligned} g_{K-1} &= a_{K-1} \\ g_k &= a_k \oplus a_{k+1}, \quad 0 \le k \le K-2 \end{aligned} \qquad (2)$$

where $\oplus$ denotes the Boolean XOR operation and a_k is the *k*-th bit representation. This Gray code representation has the unique property that consecutive codewords differ only in one bit position [12]. After this transformation, TGCBPM and WTGCBPM use respective number of non-matching points (NNMP) as matching criteria which are given as:

$$\begin{aligned} &NNMP_{TGCBPM,NTB}(m,n) \\ &= \sum_{i=0}^{N-1}\sum_{j=0}^{N-1}\sum_{k=NTB}^{K-1} 2^{k-NTB}\{g_k^t(i,j) \oplus g_k^{t-1}(i+m,j+n)\} \end{aligned} \qquad (3)$$

$$\begin{aligned} &NNMP_{WTGCBPM,NTB}(m,n) \\ &= \sum_{i=0}^{N-1}\sum_{j=0}^{N-1}\sum_{k=NTB}^{K-1} g_k^t(i,j) \oplus g_k^{t-1}(i+m,j+n) \end{aligned} \qquad (4)$$

where *K* represents the pixel-depth and *NTB* is the number of truncated bits, the motion block size is *N*×*N,* and $-s \le m, n \le s$ is the search range. The *k*th most significant bit of the Gray-coded image pixel frame of time *t* is represented as g_k^t (*i,j*). Note the similarity between $NNMP_{TGCBPM}$ and $NNMP_{WTGCBPM}$. Compared with the previous BPM based ME algorithms, TGCBPM and WTGCBPM based ME show significant gains in terms of ME accuracy [7][8].

B. Multiple Candidate Motion Search

In [11], the authors proposed multiple candidate motion searches based on two different matching criteria, i.e., $WNNMP_{2BT}(M)$ and $WNNMP_{2BT}(L)$. In general, performing many motion searches using different matching criteria simultaneously helps improving the overall performance, but their computational complexities would increase heavily with the number of matching criteria used. However, they exploited the almost identical operations between two different matching criteria and did the dual motion searches with negligible computational complexity increase. The multiple candidate motion searches based on the weighted 2BTs (MCW2BT) can be summarized as follows [11]:

1) Find two best motion vectors using two different matching criteria, respectively.
2) If two best motion vectors with the respective matching criteria are the same, declare it as the best motion vector for the current block and go to 4).
3) Calculate SADs of the two best motion vectors and declare the motion vector with less SAD as the best motion vector for the current block.
4) Go to the next current block.

It should be noted that the calculations of SADs are needed only when the two best motion vectors are different. According to the results in [11], the SAD calculations are needed about 1 out of 11 motion blocks on average which is very small. To enhance the overall ME accuracy, they also adopted the local search strategy into the MCW2BT and greatly enhanced the ME accuracy with small complexity increase.

III. PROPOSED ALGORITHMS

This section explains our proposed algorithms, which are based on Gray-coded BPMs. Since the proposed algorithms are basically extensions of the MCW2BT, in order to apply the multiple candidate motion searches, the following two conditions must be satisfied:

i) The relative local behaviors of ME results for respective matching criteria vary substantially.
ii) The operations among different matching criteria must share many identical operations.

Condition i) is for enhancing the ME accuracy. The multiple candidate motion searches take only the advantages of the respective ME results. Therefore if the local behaviors of ME results for respective matching criteria are almost the same or if a certain matching criterion based ME always outperforms the other matching criteria based ME, there is no room for improving the ME accuracy using multiple candidate motion searches. The greater the difference of ME results among different matching criteria, the better the final ME results of the multiple candidate motion searches. Condition ii) is needed for computing the different matching criteria efficiently with significantly less computations. If there are some identical operations among different matching criteria, respective matching criteria can be easily calculated using already calculated values. If this condition is not met, it is not useful for practical purposes.

To check the condition i) for the Gray-coded BPMs, we analyzed the performances of TGCBPM and WTGCBPM. Table 1 shows the average peak signal to noise ratio (PSNR) performances of TGCBPM and WTGCBPM of some CIF-size test sequences with various NTBs when the motion block size is 16×16, the search range is ±16. We can see from the *K*th column of Table I that the average performance varies from sequence to sequence. For sequences of "akiyo", "foreman", "hall", and "table tennis", WTGCBPM outperforms TGCBPM and for the other sequences, TGCBPM outperforms WTGCBPM.

Unlike MCW2BT, we can also check the condition i) for TGCBPMs and WTGCBPMs with different NTBs. That is, the local behaviors of the ME results for TGCBPM and WTGCBPM with *NTB = k* and with *NTB = l* (*k ≠ l*) are very different. From the *G*th and *H*th column of Table I, the average performances of TGCBPM with *NTB* = 5 are

TABLE I. AVERAGE SEARCH POINTS OF ALGORITHMS FOR CIF SEQUENCES WHEN THE MOTION BLOCK SIZE IS 16×16 (100-FRAME, SEARCH RANGE IS ±16)

	TGCBPM			**WTGCBPM**			**PSNR Difference**				
	A	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**	**K**
	***NTB* = 6**	***NTB* = 5**	***NTB* = 4**	***NTB* = 6**	***NTB* = 5**	***NTB* = 4**	**ΔPSNR (B-A)**	**ΔPSNR (C-B)**	**ΔPSNR (E-D)**	**ΔPSNR (F-E)**	**ΔPSNR (B-E)**
football	23.41	23.55	23.26	23.39	23.68	23.68	0.14	-0.29	0.29	0.00	0.13
akiyo	42.03	42.56	42.62	41.82	42.37	42.49	0.53	0.06	0.55	0.12	-0.19
foreman	31.91	32.68	32.83	31.76	32.51	32.82	0.77	0.15	0.75	0.31	-0.17
hall	33.09	33.84	33.71	33.01	33.81	33.79	0.75	-0.13	0.80	-0.02	-0.03
bus	24.46	24.53	24.39	24.51	24.66	24.67	0.07	-0.14	0.15	0.01	0.13
tempete	27.37	27.46	27.42	27.37	27.51	27.52	0.09	-0.04	0.14	0.01	0.05
table tennis	28.11	28.37	28.30	27.93	28.18	28.29	0.26	-0.07	0.25	0.11	-0.19
children	28.89	28.68	28.46	28.91	28.93	28.90	-0.21	-0.22	0.02	-0.03	0.25
average	**29.91**	**30.21**	**30.12**	**29.84**	**30.21**	**30.27**	**0.30**	**-0.09**	**0.37**	**0.06**	**0.00**

always better, but between *NTB* = 4 and *NTB* = 5, the average performances vary slightly from sequence to sequence. However, we can observe the alternating local ME results of TGCBPM with *NTB* = 5 and *NTB* = 6 in Figure 1. Figure 1 shows the frame level PSNR results of sequence "container" using TGCBPM with *NTB* = 5 and *NTB* = 6. For the matching criterion of WTGCBPM, we can see the alternating PSNR performances very clearly in the *I*th and *H*th column of Table I.

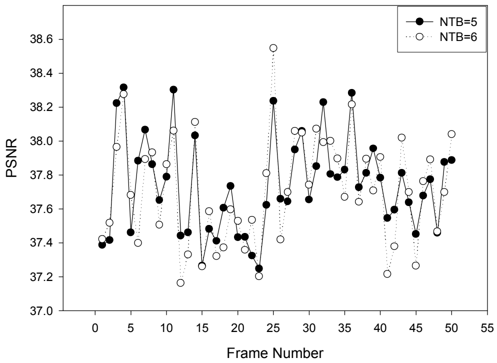


Figure 1. Frame-wise PSNR results of "container" using TGCBPM with *NTB* = 5 and *NTB* = 6 when the motion block size is 16×16 and the search range is ±16.

For checking the condition ii) for the above discussed, we define the following $NNMP_{gram,k}$ of the *k*th most significant bit as:

$$NNMP_{gram,k}(m,n) := \sum_{i=0}^{N-1}\sum_{j=0}^{N-1} g_k^t(i,j) \oplus g_k^{t-1}(i+m,j+n) \qquad (5)$$

Then, the matching error criteria of TGCBPM and WTGCBPM can be compactly represented as

$$\begin{aligned} NNMP_{TGCBPM,NTB}(m,n) &= \sum_{k=NTB}^{K-1} 2^{k-NTB} NNMP_{gram,k}(m,n) \\ NNMP_{WTGCBPM,NTB}(m,n) &= \sum_{k=NTB}^{K-1} NNMP_{gram,k}(m,n) \end{aligned} \qquad (6)$$

Therefore, TGCBPM and WTGCBPM with different NTBs can be recurrently calculated as follows:

$$\begin{aligned} & NNMP_{TGCBPM,K-1}(m,n) = NNMP_{gram,K-1}(m,n) \\ & NNMP_{TGCBPM,k}(m,n) \\ & = 2 \times NNMP_{TGCBPM,k+1}(m,n) + NNMP_{gram,k}(m,n) \end{aligned} \qquad (7)$$

$$\begin{aligned} & NNMP_{WTGCBPM,K-1}(m,n) = NNMP_{gram,K-1}(m,n) \\ & NNMP_{WTGCBPM,k}(m,n) \\ & = NNMP_{WTGCBPM,k+1}(m,n) + NNMP_{gram,k}(m,n) \end{aligned} \qquad (8)$$

where (0 ≤ *k* ≤ *K*-2). Note that in this case, the recurrence relation is in reverse order. From the equations (6), (7) and (8), we can see that once the calculations of the values $NNMP_{gram,k}$ are finished, the final calculations of the matching criteria $NNMP_{TGCBPM,K\text{-}1}$, $NNMP_{TGCBPM,K\text{-}2}$, ···, and $NNMP_{TGCBPM,NTB}$ can be easily carried out with additional (*K-NTB*-1) additions and (*K-NTB*-1) shift operations only. In the same way, for calculations of the matching criteria $NNMP_{WTGCBPM,K\text{-}1}$, $NNMP_{WTGCBPM,K\text{-}2}$, ···, and $NNMP_{WTGCBPM,NTB}$ can be easily calculated with additional (*K-NTB*-1) additions only. Therefore, we can enhance the ME accuracy using multiple candidate motion searches without noticeable increase of computational complexity. Exploiting the above observations, we propose a multiple candidate motion searches based on the Gray-coded BPM (MCGCBPM) as follows:

1) Calculate the values of $NNMP_{gram,k}$ ($NTB \le k \le K$-1) as in (5).
2) Using (7), calculate the matching criteria of TGCBPM, i.e., $NNMP_{TGCBPM,K\text{-}1}$, $NNMP_{TGCBPM,K\text{-}2}$, ···,

TABLE II. AVERAGE PSNR AND THE NUMBER OF SAD CALCULATIONS COMPARISON OF THE ALGORITHMS ($NTB = 4$)

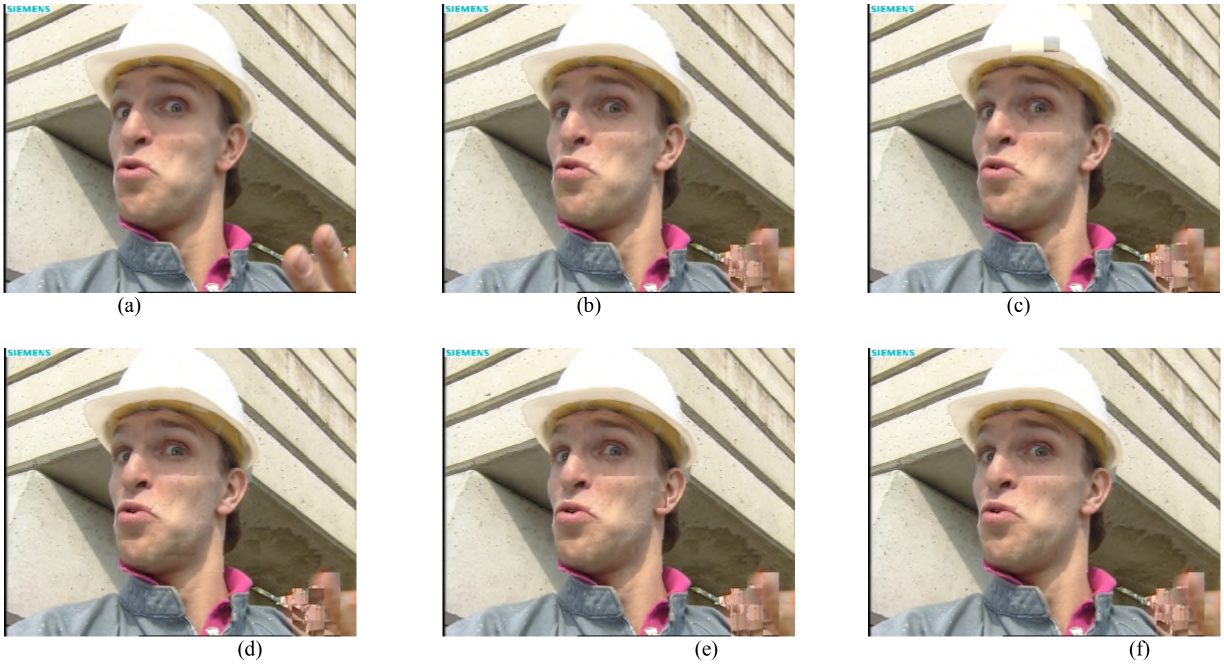
	1BT		2BT		TGCBPM		WTGCBPM		AM2BT		MCW2BT -LS		MCGCBPM -LS		FSBMA	
sequences	16×16	8×8	16×16	8×8	16×16	8×8	16×16	8×8	16×16	8×8	16×16	8×8	16×16	8×8	16×16	8×8
stefan	25.12	24.90	25.25	25.77	25.48	26.28	25.40	26.16	25.68 (23.95)	26.61 (13.80)	25.61 (10.65)	26.55 (12.11)	25.70 (6.63)	26.68 (7.46)	25.75	26.74
football	22.64	23.46	23.06	24.35	23.68	25.36	23.26	24.90	23.96 (56.76)	25.78 (16.13)	23.82 (11.90)	25.55 (12.18)	23.94 (7.32)	25.86 (7.40)	24.00	25.95
akiyo	41.66	35.23	42.43	41.81	42.49	42.66	42.62	42.93	42.57 (0.57)	42.61 (0.63)	42.79 (3.61)	43.07 (4.47)	42.83 (2.49)	43.44 (2.91)	42.84	43.48
foreman	31.69	30.64	31.81	31.98	32.82	33.88	32.83	34.06	32.60 (3.26)	33.66 (1.95)	33.02 (12.17)	34.11 (13.72)	33.38 (7.31)	34.99 (8.11)	33.43	35.13
mobile	23.50	23.31	23.58	23.89	23.78	24.44	23.58	24.05	23.84 (37.24)	24.6 (15.79)	23.89 (9.93)	24.71 (11.73)	23.91 (6.10)	24.79 (6.98)	23.92	24.83
hall	32.13	30.83	33.22	33.95	33.79	34.94	33.71	34.89	33.81 (2.82)	34.97 (1.29)	33.93 (11.77)	35.24 (13.68)	34.27 (7.45)	35.73 (8.46)	34.34	35.87
coastguard	29.09	28.17	29.23	29.52	29.44	30.09	29.37	30.25	29.55 (7.85)	30.36 (4.77)	29.56 (9.92)	30.41 (11.42)	29.61 (6.37)	30.64 (7.13)	29.62	30.68
container	37.57	34.81	38.12	37.82	37.90	37.69	37.98	37.84	38.16 (0.59)	38.09 (0.67)	38.24 (7.91)	38.26 (9.82)	38.24 (5.09)	38.26 (6.05)	38.33	38.42
bus	23.86	24.58	24.14	25.24	24.67	26.33	24.39	25.97	24.79 (35.79)	26.52 (11.76)	24.64 (10.84)	26.34 (11.91)	24.84 (6.72)	26.74 (7.17)	24.90	26.83
dancer	29.67	30.57	30.29	30.87	31.55	32.64	30.93	32.21	31.45 (11.37)	32.66 (4.18)	31.45 (17.55)	32.56 (17.66)	31.92 (10.74)	33.40 (10.67)	32.14	33.72
mother and daughter	37.58	35.83	39.34	38.99	39.57	39.59	39.75	40.08	39.58 (1.18)	39.67 (0.85)	39.84 (11.73)	40.36 (13.38)	40.06 (7.14)	40.86 (7.94)	40.12	41.02
tempete	27.01	26.51	27.25	27.46	27.52	27.96	27.42	27.88	27.63 (19.10)	28.16 (8.45)	27.60 (9.87)	28.18 (11.64)	27.68 (6.17)	28.33 (7.02)	27.70	28.38
table tennis	27.44	28.09	27.87	29.00	28.29	29.83	28.3	29.8	28.65 (10.61)	30.10 (4.08)	28.62 (11.79)	30.11 (12.77)	28.77 (7.16)	30.44 (7.59)	28.87	30.55
flower	25.73	26.44	25.83	26.92	25.95	27.23	25.88	27.15	25.97 (22.36)	27.30 (9.05)	26.00 (11.38)	27.37 (12.18)	26.02 (6.90)	27.43 (7.33)	26.03	27.45
children	28.05	27.75	28.32	29.64	28.90	30.55	28.46	30.16	29.16 (17.64)	30.92 (5.29)	29.08 (4.59)	30.76 (4.33)	29.19 (3.01)	31.02 (2.80)	29.24	31.10
paris	30.16	29.73	30.16	31.23	30.53	31.96	30.36	31.75	30.58 (7.33)	32.09 (2.92)	30.62 (6.91)	32.19 (8.56)	30.69 (4.27)	32.36 (5.24)	30.71	32.41
news	35.58	33.22	36.50	37.20	37.05	38.19	36.82	38.11	37.05 (3.05)	38.32 (1.38)	37.18 (5.53)	38.50 (6.64)	37.32 (3.58)	38.89 (4.16)	37.33	38.95
Average	**29.91**	**29.06**	**30.38**	**30.92**	**30.79**	**31.74**	**30.65**	**31.66**	**30.88** **(15.38)**	**31.91** **(6.06)**	**30.93** **(9.87)**	**32.02** **(11.07)**	**31.08** **(6.14)**	**32.34** **(6.73)**	**31.13**	**32.44**

and $NNMP_{TGCBPM,NTB}$. And using (8), calculate the matching criteria of WTGCBPM, i.e., $NNMP_{WTGCBPM,K\text{-}1}$, $NNMP_{WTGCBPM,K\text{-}2}$, ···, and $NNMP_{WTGCBPM,NTB}$.

3) Find respective candidate motion vectors according to the matching criteria calculated in 2).
4) Calculate SADs of candidate motion vectors and declare the motion vector with the least SAD as the best motion vector for the current block.
5) Go to the next current block.

To enhance the overall ME accuracy, we adopt the local search strategy into MCGCBPM. As we tested several local search algorithms with MCGCBPM, two-step search in [10] showed the best performance in terms of the PSNR performance and computational complexity. In summary, we propose the following MCGCBPM with two-step local search, namely MCGCBPM-LS:

1) Calculate the values of $NNMP_{gram,k}$ ($NTB \leq k \leq K$-1) as in (5).
2) Using (7), calculate the matching criteria of TGCBPM, i.e., $NNMP_{TGCBPM,K\text{-}1}$, $NNMP_{TGCBPM,K\text{-}2}$, ···, and $NNMP_{TGCBPM,NTB}$. And using (8), calculate the matching criteria of WTGCBPM, i.e., $NNMP_{WTGCBPM,K\text{-}1}$, $NNMP_{WTGCBPM,K\text{-}2}$, ···, and $NNMP_{WTGCBPM,NTB}$.
3) Find respective candidate motion vectors according to the matching criteria calculated in 2).
4) Calculate SADs of candidate motion vectors and declare the motion vector with the least SAD as the predicted motion vector.
5) Do two-step search around the predicted motion vector and find the best motion vector using SAD for the current block.
6) Go to the next current block.



(a) (b) (c)

(d) (e) (f)

Figure 2. Sample results for the “foreman” sequence when the motion block size is 16 × 16 and the search range is ±16. (a) Original frame, (b) Reconstructed frame with the SAD based ME (PSNR = 30.48dB) (c) Reconstructed frame with the MCW2BT-LS ME (PSNR = 29.56dB), (d) Reconstructed frame with TGCBPM with NTB = 5 (PSNR = 29.91dB), (e) Reconstructed frame with WTGCBPM with NTB = 5 (PSNR = 29.72dB), and (f) Reconstructed frame with the proposed MCGCBPM-LS with NTB = 5 (PSNR = 30.36dB)

IV. Experimental Results

Several experiments were carried out to compare the performances of the proposed algorithms with 1BT [3], 2BT [4], TGCBPM [7], WTGCBPM [8], AM2BT [10], MCW2BT-LS [11], and the FSBMA. The first 100 frames of 17 CIF (352 × 288) sequences (stefan, football, akiyo, foreman, mobile, hall monitor, coastguard, container, bus, dancer, mother and daughter, tempete, table tennis, flower, children, paris, and news) are used as test sequences. All the searching processes were in spiral order. For an efficient calculation of SADs, we adopted a typical PDE algorithm whose partial SAD order is 8.

A. *Performance Evaluation*

Table II shows the average PSNR results and the number of SAD calculations (if available) per motion block of the proposed MCGCBPM-LS compared with other ME algorithms when NTB = 4. When the motion block size is 16×16 and the search range is ±16, we can see that the proposed MCGCBPM -LS outperforms 1BT and 2BT based ME by 1.17dB and 0.70dB, respectively. And since the proposed MCGCBPM-LS takes only the advantages of TGCBPM and WTGCBPM, we can expect that the performance of the proposed algorithm will always be better than TGCBPM and WTGCBPM based ME, and the table shows that it is the case. Compared with AM2BT and MCW2BT-LS, the proposed MCGCBPM-LS is better in terms of both the PSNR performance and the computational complexity. To be specific, for PSNR performance, the proposed algorithm is better than AM2BT and MCW2BT-LS by 0.20dB and 0.15dB, respectively. For computational complexity in terms of SAD computations, the proposed MCGCBPM-LS requires 39.92% of AM2BT and 62.21% of MCW2BT-LS. Compared with the performance with FSBMA, FSBMA outperforms the proposed MCGCBPM-LS by 0.05dB, which is very small. To be specific, for sequences “akiyo”, “mobile”, “coastguard”, “tempete”, “flower”, “paris”, and “news”, the PSNR performance differences are within 0.03dB.

When the motion block size is 8×8 and the search range is ±8, the PSNR performance differences between the proposed MCGCBPM-LS and 1BT, 2BT, TCGBPM, WTGCBPM, AM2BT, and MCW2BT-LS are more than twice compared with the results when the motion block size is 16×16. For computational complexity, the proposed MCGCBPM-LS requires 60.79% of MCW2BT-LS in terms of SAD computations. In contrast, the proposed MCGCBPM -LS requires 11.06% more SAD computations compared with AM2BT. However, since the PSNR improvement of the proposed MCGCBPM-LS over AM2BT is about 0.43dB, we presume that this slight computational complexity increase is tolerable.

B. *Visual Quality Evaluation*

To compare the visual quality of the proposed algorithm with other algorithms (TGCBPM, WTGCBPM, MCW2BT-LS, and FSBMA), the reconstructed frames of “foreman” are given in Figure 2. The motion block size is 16×16, the search range is ±16, and NTB = 5. As can be seen from the figures, many bad motion vectors are observed in the reconstructed frames of TGCBPM and WTGCBPM resulting in annoying visual quality. For MCW2BT-LS, some bad motion vectors are seen in the reconstructed frames. However, in case of the proposed MCGCBPM-LS, bad motion vectors are reduced significantly and its visual appearance is much more pleasing and almost the same as that of the FSBMA, which can be expected from the PSNR results.

V. Conclusions

Low complexity ME algorithms based on the Gray-coded BPM are proposed in this paper. By exploiting almost identical operations among similar but different matching error criteria, we can efficiently determine the respective candidate motion vectors. Moreover, adopting multiple candidate ME strategies into those candidate motion vectors and two-step local searches around the best candidate motion vector enhance the motion estimation accuracy substantially with relatively small computational complexity increase. Experiments were carried out for comparing the performance of the proposed algorithms with other BPM based motion estimation algorithms, and FSBMA as well. The proposed MCGCBPM -LS outperforms all the other BPM based MEs with negligible complexity increase. The PSNR difference between the proposed MCGCBPM-LS and FSBMA is within 0.05dB on average for *NTB* = 4.

References

[1] Information Technology - Coding of Audio Visual Objects - Part 2: Visual, JTC1/SC29/WG11, ISO/IEC 14496-2 (MPEG-4 Visual), 2002.

[2] Advanced Video Coding for Generic Audiovisual Services, ITU-T Recommendation H.264, May 2005.

[3] B. Natarajan, V. Bhaskaran, and K. Konstantinides, “Low-Complexity Block-based Motion Estimation via One-Bit Transforms,” IEEE Trans. Circuits and Systems for Video Tech., vol. 7, no. 5, Aug. 1997, pp. 702-706.

[4] S. Erturk, “Multiplication-Free One-Bit Transform for Low-Complexity Block-Based Motion Estimation,” IEEE Signal Processing Letters, vol. 14, no. 2, Feb. 2007, pp. 109-112.

[5] O. Urhan and S. Erturk, “Constrained One-Bit Transform for Low Complexity Block Motion Estimation,” IEEE Trans. Circuits and Systems for Video Tech., Apr. 2007, vol. 17, no. 4, pp. 478-482.

[6] A. Erturk and S.Erturk, “Two-Bit Transform for Binary Block Motion Estimation,” IEEE Trans. Circuits and Systems for Video Tech., vol. 15, no. 7, Jul. 2005, pp. 938-946.

[7] A. Celebi, O. Akbulut, O. Urhan, and S. Erturk, “Truncated Gray-Coded Bit-Plane Matching Based Motion Estimation Method and Its Hardware Architecture,” IEEE Trans. Consumer Electron., vol. 55, no. 3, Aug. 2009. pp. 1530-1536.

[8] A. Celebi, H. Lee, and S. Erturk, “Bit Plane Matching Based Variable Block Size Motion Estimation Method and Its Hardware Architecture,” IEEE Trans. Consumer Electron., vol. 56, no. 3, Aug. 2010, pp. 1625-1633.

[9] C. Choi and J. Jeong, “Enhanced Two-bit Transform based Motion Estimation via Extension of Matching Criterion,” IEEE Trans. Consumer Electron, vol. 56, no. 3, Aug. 2010, pp. 1883-1889.

[10] B. Demir and S. Erturk, “Block Motion Estimation Using Adaptive Modified Two-Bit Transform,” IET Image Processing, vol. 1, no. 2, Jun. 2007, pp. 215-222.

[11] C. Choi and J. Jeong, “Low complexity Weighted Two-bit Transforms based multiple candidate Motion Estimation,” IEEE Trans. Consumer Electron, vol. 57, no. 4, Nov. 2011, pp. 1837-1842.

[12] S. Ko, S. Lee, S. Jeon, and E. Kang, “Fast Digital Image Stabilizer based on Gray-coded Bit-plane Matching,” IEEE Trans. Consumer Electron., vol. 45, no. 3, Aug. 1999, pp. 598-603.

Towards Audio Enrichment through Images: A User Evaluation on Image Relevance with Spoken Content

Danish Nadeem
Human Media Interaction
University of Twente
Enschede, Netherlands
Email: d.nadeem@utwente.nl

Mariët Theune
Human Media Interaction
University of Twente
Enschede, Netherlands
Email: m.theune@utwente.nl

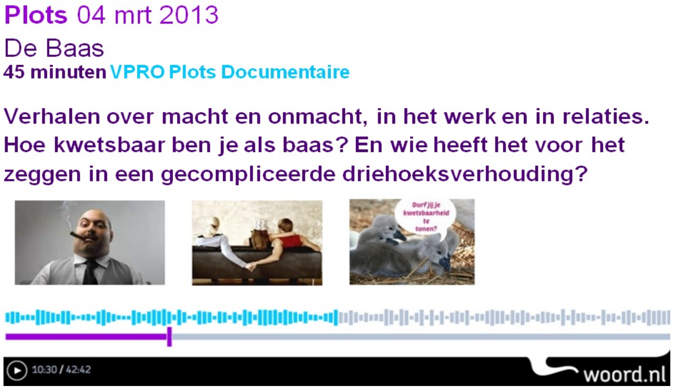
Roeland Ordelman
Human Media Interaction and
Netherlands Institute for Sound and Vision
Hilversum, Netherlands
Email: rordelman@beeldengeluid.nl

***Abstract*—In a visual radio scenario, where radio broadcast is consumed on mobile devices (such as phones and tablets), watching pictures as you listen, may improve information or entertainment value of the programme. We assume that audio enrichment through images can be useful to users when the selection of images is semantically associated to the spoken content. In this paper, we report about a user study to evaluate the relevance of images selected automatically based on the speech content of audio fragments (audio interviews in the Dutch language). A total of 43 participants took part in the study. They listened to a set of audio fragments and performed an image rating task. In addition to that, we conducted a small follow-up study with 3 participants to shed more light on the results of the first study. We observed that merely keyword similarity between image captions and speech fragments may not be a good predictor for image relevance from a user viewpoint, and therefore we speculate that taking topic of speech into account may improve image relevance. Furthermore, from a user perspective on the added value of audio enrichment with images, we learned that the images should strengthen the understanding of audio content rather than distracting the listeners. The insights gained in the study will open room for further investigation of audio enrichment through images and its effect on user experience.**

***Keywords*–user evaluation study; linking audiovisual archives; multimedia semantics; audio augmentation.**

I. Introduction

In recent times, as multimedia content has become increasingly easy to produce and process, technologies are being developed to automatically enrich media with links to additional information and create applications which access the multimedia content. The idea is to provide new added value services to consumers for information, education and entertainment [1]. Advances in audiovisual content enrichment techniques have generated interest in various domains like class lectures [2] and meetings [3] [4]. One of the applications - especially for audio enrichment - is the *visual radio* application, where the idea is to complement radio programmes with additional information in various modalities (e.g., text, images and videos) automatically. In this paper, we focus on enrichment of audio programmes like radio interviews, by presenting topically related images (i.e., images that are somewhat semantically associated with the speech), drawn automatically from an image collection (see an example application in Figure 1). The images presented in the figure are selected on the basis of the similarity between the keywords used to describe the image and keywords in the transcript of the audio fragment. For example, the topic of the audio fragment can be determined by a set of keywords like *baas* (Boss), *driehoeksverhouding* (love triangle) and *kwetsbaar* (vulnerable). Images containing these keywords in their description (caption, meta-data or annotation) are selected as relevant images.


Plots 04 mrt 2013
De Baas
45 minuten VPRO Plots Documentaire

Verhalen over macht en onmacht, in het werk en in relaties. Hoe kwetsbaar ben je als baas? En wie heeft het voor het zeggen in een gecompliceerde driehoeksverhouding?

Figure 1. A possible audio player interface showing images that are annotated with the terms also occurring in the spoken content. Interface source: Dutch Public Radio Broadcast platform (woord.nl).

Audio enrichment through images may deliver a richer experience for entertainment and provide additional visual information on spoken content for listeners. For the selection of images related to a speech fragment, typically a string-matching approach is taken by comparing the content of the speech transcript to the textual information of the image such as keywords, caption and meta-data, etc., used to describe the image. Generally, transcripts from speech are analyzed to extract knowledge from the speech in the form of named entities [5]. Then these named entities are used as search terms to find images from an image database collection. The relevance of the image is determined by measuring the similarity between the textual representation of the image and the search terms using a document retrieval-based approach [6]. However, such approaches developed for retrieval do not take into account the other aspects surrounding the speech such as time-frame, situations and topics. As a consequence they may not optimally represent what is said in the audio programme. Furthermore, context may play a crucial role to improve entertainment value by enriching audio content with related images, since each listener has different interests, values and preferences.

To gain some insight on the user perspective, and to understand image relevance for the presentation of images with audio, we seek answers to the following questions:

1) Do users perceive (automatically selected) images as an added value, when presented with audio content?
2) Is there any good predictor for the relevance of images for a given speech content?

We hypothesize that for the decision of presenting an image to a user, the content of the image should match with the topic of the speech content in order to add value to visualization. For example, if the speech is about "how-to" make *"Italian Tiramisu cake"*, whereas an image presented to a user shows an *"Italian Pizza"*, it will confuse the user because the topic is not fully matched. To test our hypothesis, we conducted evaluation studies, where we asked users to assess the relevance of images with the speech content of audio fragments. We consider this as a first step to develop our understanding of audio enrichment with images from a user perspective.

In the following sections, we will discuss related work on enrichment of various media (Section II), then we will describe our evaluation studies and discuss their respective results (Sections III and IV) and finally draw conclusions concerning our questions (Section V).

II. RELATED WORK

Presenting images in audio programmes is an instantiation of what is generally called *semantic linking*, and which has gained a lot of attention recently in the audiovisual content retrieval and linking research community [7] [8] [9]. There have been investigations about hyperlinking from text sources [10] [11]. Among the latest work is semantic linking of Twitter posts with related news articles to contextualize Twitter activities [12]. Text enrichment through linking of images has found useful applications in multimodal question answering systems [13] and learning scenarios [14]. Recent research towards understanding the user perspective in image enrichment [15] and audio enrichment is also reported in our previous work [16]. Related work on video enrichment by linking additional resources through semi-automatic methods for a news show broadcast scenario is reported in the context of the LinkedTV project [17]. In the direction of audio enrichment, there have been studies on audiovisual chat conversation enrichment by linking Flickr (www.flickr.com) images to the topic of conversation [18]. Furthermore, with the emergence of visual radio applications, various techniques are deployed such as allowing users to tag contents (comments or images) on the audio timeline through an interactive audio player [19].

Towards audio enrichment, we intend to expand our general understanding of image relevance with speech content in audio programmes. Here, we focus on automatically linking speech content to related images - where a link connects an 'anchor' (information source) to a 'target' (information destination). In a speech to images linking situation, we consider an anchor as a spoken word or a phrase in an audio programme, such as the name of a person, a topic, a place, an event, location, etc., while a target can be topically related image drawn from an image database. In practice, multiple links may be created from an anchor to different target images.

III. IMAGE EVALUATION USER STUDY

We conducted an image evaluation user study where we asked participants to listen to audio fragments and provided them with a set of (automatically selected) images from a Dutch National Archive collection (www.gahetna.nl). Participants were asked to rate the suitability of the images according to the information they heard in the audio fragments. In the following section we describe the method of our study.

A. Participants

Forty three native Dutch people who were all able to understand written and spoken Dutch participated in an online user study. Some of the participants were colleagues and common friends but most of the participants were from the general public, whom we found by visiting a public library in the town. Some of them said that they frequently listen (2-3 times a week) to radio programmes via public radio broadcast. We asked for their consent and emails to participate in the study. Later, we sent them an email with a survey link to participate in the study. Participants were between 25 and 67 years old (M = 44.7, SD = 16.2). Of the forty three participants, twenty four were female and nineteen were male.

B. Materials

For the 43 participants who participated in the image evaluation study, we randomly selected four short-duration audio fragments (ranging from 2 to 5 minutes duration), from a collection of marathon audio interviews. The interviews in the audio were spoken in the Dutch language. The decision to use short-duration fragments was taken for practical reasons to keep study duration limited to 30 minutes. The audio interviews are publicly available at the Dutch Public Radio Broadcast platform. Furthermore, an image database collection of the Dutch National Archive, containing over 14 million historical images is used to find images relevant to Dutch audio fragments. A total of 40 images from the collection were used in the study. To draw images for a speech transcript, the images were indexed using Lucene plugin (Apache Lucene^{TM}), based on the keyword in captions and image descriptions. The search were performed using Elasticsearch® to retrieve images whose keyword meta-data match with the keyword in the transcript of an audio fragment. Because the audio fragments varied in speech content and duration (from 2 to 5 minutes), each fragment was presented with a different number of images.

C. Rating task

All of the participants were asked to listen to each of the 4 audio fragments in the same order. We informed the participants that they could listen to the audio as many times as they liked before moving to the next audio fragment. After listening to each fragment, they had to rate on a 5-point likert scale how familiar they were with the spoken content of the audio fragment. Subsequently, for every fragment participants were presented with a varying number of images together with their captions. For fragment 1, they were presented with 14 images. For fragment 2, they were presented with 4 images. For fragment 3, they were presented with 8 images and for fragment 4, they were presented with 14 images, making a total of 40 images for all 4 audio fragments. Furthermore, the experiment interface was designed such that the participants could listen to the fragment by clicking a play button on top of the page, and then scroll down the page to see the images they were asked to rate. The images were presented statically (all at a time) according to the order in which speech was delivered

TABLE I. RELEVANCE SCORES FOR ALL THE IMAGES ACCORDING TO USER RATINGS.

Relevance score	% of user ratings
strongly disagree	50%
disagree	40%
neither agree nor disagree	5%
agree	2.9%
strongly agree	2.1%

in the audio fragment. For example, suppose if the speech fragment says: "Obama visits Paris and meets the president Hollande", the image of Obama is placed before the image of Hollande in the web interface.

For each image we asked the participants to provide a rating, indicating to which extent they agree that there is a match between the image and the speech fragment. The rating was distributed on a 5-point likert scale of agreement (1 = strongly disagree; 2 = disagree; 3 = neither agree nor disagree; 4 = agree; and 5 = strongly agree).

D. Additional survey questionnaire

After the participants completed the image rating task for all of the four audio fragments, we asked them to respond to an additional survey questionnaire. The questionnaire consisted of two statements to be rated on the same 5-point likert scale for agreement, and two general questions on how frequently they listened to Internet radio and if they used mobile devices (phones, tablets, etc.,) for listening to Internet radio. The statements we asked them to rate were (translated from Dutch): (i) *The content of the image should match with the topic of the fragment.* (ii) *The images are still useful, even when the content of the images does not match with the content of the speech.* Finally, we also provided an option for general feedback to the participants about their perspective on combining audio programmes with images.

E. Results of user responses

We analysed user rating responses from the likert-scale as a total percentage of participants' agreement with the relevance of images for the speech fragment. The result is presented in the Table I. From the result, we found that the overall image relevance scores were very low. This suggests that the selection algorithm did not perform well to select images that user would find relevant. Furthermore, to know the agreement among the participants, we computed inter-user agreements using Krippendorff's alpha [20], which was calculated using the SPSS software and a macro. We found the value of Krippendorff's alpha (α) = 0.52, which is considered "fair agreement" level.

For the results on the additional survey questionnaire, 88% participants agreed that the content of the image should match with the topic of the fragment. Whereas, only 6% agreed that they may consider images useful even if the image content mismatches with the topic of speech fragment. Furthermore, concerning Internet radio: 23% listened several times per week, 6% listened 2-3 times per week, 18% listened once per week, 6% listened once per month, 12% listened less often and 35% never listened to Internet radio. Concerning the use of device for listening to audio, 50% responded that they used mobile devices such as phones or iPads to listen to the Internet radio. Furthermore, we received some feedback concerning the user perception on added value of images in audio programmes, as some of the participants gave general feedback in the survey questionnaire. Their responses suggest that generally the idea of combining audio and visual modalities is interesting. However, the images should strengthen the understanding of audio content rather than distracting the listeners.

F. Analysis of user ratings

As the results obtained from the image evaluation user study showed low image relevance scores, we analysed the caption of each image and compared it with the transcript of the speech fragment where the image was presented. This gave us an indication of whether the selected images are relevant to the speech fragment.

To analyze all the images with their captions, we listed the main keywords (for, e.g., named-entities) from the caption of an image. Then we checked if the same keyword was present in the speech fragment. If it was indeed the case, and if the image appeared to be somewhat associated with the speech fragment, we considered that image to be relevant to the fragment. For example, there was a speech fragment mentioning the keywords *"Red Indian"* (native Americans). To enrich this fragment, the algorithm selected two images with captions that had the keyword *"Indian"*. However, one of the images showed a native American (caption containing the words *"Red Indian"*), whereas the other image showed *soldiers from India during World War 2*. Looking only at the captions, both of the images may be considered relevant, however, looking at the speech topic (aboutness of the speech fragment), only the first image can be considered relevant to the speech fragment.

We observed that the images matching the topic of a speech fragment in our subjective analysis, were also the images that were given higher relevance score in the user response study. To shed light on our analysis, we conducted a small follow-up study with three participants to assess the influence of image caption. We assumed that an image without a caption may or may not convey its relevance to the content of a speech fragment. Whereas images together with their caption will provide additional content information about the images, and therefore, will be rated higher on relevance when the image caption matches with the topic of a speech fragment. The study is described in the following section.

IV. INFLUENCE OF IMAGE CAPTION STUDY

We asked the raters to provide a relevance score between an image and the speech fragment on the basis of: (i) image only and (ii) image with caption.

A. Participants

We asked three native Dutch male participants, ranging in age from 27 to 42 years old (M = 34.4, SD = 7.5) to participate in the study. None of them were part of the image evaluation user study described in the Section III.

B. Materials

For these 3 raters, we used the transcript of each of the 4 audio fragments from the first study and provided them with the same 40 images for an assessment task. We decided to use speech transcripts of the audio fragments so that the participants could clearly identify the spoken words, which might have been misheard in listening to the audio fragment.

C. Tasks

Similar to the image evaluation user study, we presented all the images to all the 3 raters in the same order for each transcript of the audio fragment. However, here the task was performed in two steps; (i) all the raters were asked to read through a transcript of each of the audio fragments and then they were asked to assess the images without caption, (ii) after they had assessed the images, they were given the same images with caption, and were asked to assess them again. The instruction given was: *"From the following images, which ones do you think are suitable for the narrative"*. Furthermore, we instructed the raters to perform a binary assessment (choosing either yes or no) for the images.

D. Results

The results of the study again show a low average relevance score for the images. Only 10.5% of the images without captions were considered relevant to the topic of the speech, when measured on a binary scale. For the image with captions, the relevance score was 17.5%. The difference between the results suggests the influence of captions in providing additional information about the images. However, because of the small sample size of the relevant images, the results remain inconclusive. Furthermore, to know the agreement among the participants, we computed the inter-rater reliability. We used Fleiss' kappa, which calculates the degree of agreement in classification over that which would be expected by chance [21]. We found the value of Fleiss' kappa (κ) = 0.64, which is considered as "substantial agreement" among the raters.

V. Conclusions

Our main observations from the evaluation studies are that the overall relevance score for images is very low. We found that merely the presence of similar keyword(s) in the image caption and the speech content is not a good predictor of image relevance. We found that only the images whose caption and visual content appeared to be related to the topic of the speech fragment were rated higher on relevance. This suggests that taking into account the topic and other aspects of speech may improve image relevance score. Furthermore, the caption seems to help people see the relevance of an image. We consider this study as a first step towards understanding audio enrichment with images from a user viewpoint. The study has given some insights on the relevance of images to speech content. We noticed that the topic of speech plays an important role for improving image relevance score. In future work, we intend to further explore the user aspects on audio enrichment and compare user experience with different modalities.

Acknowledgment

This research was supported by the Dutch national programme COMMIT (project P1-INFINITI).

References

[1] L. Nixon, "Introducing linked television: a broadcast solution for integrating the web with your tv content," in ACM International Conference on Interactive Experiences for Television and Online Video, Brussels, Belgium, June 2015, pp. 1–2.

[2] S. Mukhopadhyay and B. Smith, "Passive capture and structuring of lectures," in Proceedings of the 7th International Conference on Multimedia. New York, USA: ACM, 1999, pp. 477–487.

[3] P. Chiu, J. Foote, A. Girgensohn, and J. Boreczky, "Automatically linking multimedia meeting documents by image matching," in Proceedings of the 11th Conference on Hypertext and Hypermedia. New York, USA: ACM, 2000, pp. 244–245.

[4] A. Popescu-Belis, E. Boertjes, J. Kilgour et al., "The amida automatic content linking device: Just-in-time document retrieval in meetings," in Machine Learning for Multimodal Interaction, ser. LNCS, A. Popescu-Belis and R. Stiefelhagen, Eds., no. 5237. Springer, 2008, pp. 272–283.

[5] E. Brown, S. Srinivasan, A. Coden et al., "Towards speech as a knowledge resource," in Proceedings of the 10th International Conference on Information and Knowledge Management, ser. CIKM '01. New York, USA: ACM, 2001, pp. 526–528.

[6] P. Ferragina and U. Scaiella, "Tagme: On-the-fly annotation of short text fragments (by wikipedia entities)," in Proceedings of the 19th International Conference on Information and Knowledge Management, ser. CIKM '10. New York, USA: ACM, 2010, pp. 1625–1628.

[7] R. Mihalcea and A. Csomai, "Wikify!: Linking documents to encyclopedic knowledge," in Proceedings of the 16th Conference on Conference on Information and Knowledge Management, ser. CIKM '07. New York, USA: ACM, 2007, pp. 233–242.

[8] D. Milne and I. H. Witten, "Learning to link with wikipedia," in Proceedings of the 17th Conference on Information and Knowledge Management, ser. CIKM '08, New York, USA, 2008, pp. 509–518.

[9] D. Odijk, E. Meij, and M. de Rijke, "Feeding the second screen: semantic linking based on subtitles," in OAIR, 2013, pp. 9–16.

[10] S. Chakrabarti, B. Dom, D. Gibson et al., "Automatic resource compilation by analyzing hyperlink structure and associated text," in Proceedings of the 7th WWW conference, 30 (1-7), 1998, pp. 65–74.

[11] C. Y. Wei, M. B. Evans, M. Eliot et al., "Influencing web-browsing behavior with intriguing and informative hyperlink wording." J. Information Science, vol. 31, no. 5, 2005, pp. 433–445.

[12] F. Abel, Q. Gao, G.-J. Houben, and K. Tao, "Semantic enrichment of twitter posts for user profile construction on the social web," in Proceedings of the 8th Extended Semantic Web Conference: Research and Applications, ser. ESWC'11. Springer-Verlag, 2011, pp. 375–389.

[13] W. Bosma, "Image retrieval supports multimedia authoring," in Linguistic Engineering meets Cognitive Engineering in Multimodal Systems, E. Zudilova-Seinstra and T. Adriaansen, Eds., 2005, pp. 89–94.

[14] R. Mihalcea and C. W. Leong, "Toward communicating simple sentences using pictorial representations." Machine Translation, vol. 22, no. 3, 2008, pp. 153–173.

[15] R. Aly, K. McGuinness, M. Kleppe et al., "Link anchors in images: Is there truth?" in Proceedings of the 12th Dutch Belgian Information Retrieval Workshop, Ghent, Belgium, 2012, pp. 1–4.

[16] D. Nadeem, R. Ordelman, R. Aly, and F. de Jong, "User perspectives on semantic linking in the audio domain," in 10th International Conference on Signal-Image Technology and Internet-Based Systems, SITIS 2014. IEEE Computer Society, November 2014, pp. 244–247.

[17] D. Stein, E. Apostolidis, V. Mezaris et al., "Enrichment of news show videos with multimodal semi-automatic analysis," in Networked and Electronic Media, Istanbul, Turkey, October 2012, pp. 1–6.

[18] J. Vanattenhoven, C. van Nimwegen, M. Strobbe, O. Van Laere, and B. Dhoedt, "Enriching audio-visual chat with conversation-based image retrieval and display," in Multimedia. New York, USA: ACM, 2010, pp. 1051–1054.

[19] D. Schuurman, L. D. Marez, and T. Evens, "Content for mobile television: Issues regarding a new mass medium within today's ict environment." in Mobile TV: Content and Experience, ser. HCI, A. Marcus, A. C. Roibás, and R. Sala, Eds. Springer, 2010, pp. 143–163.

[20] A. F. Hayes and K. Krippendorff, "Answering the call for a standard reliability measure for coding data," Communication Methods and Measures, vol. 1, no. 1, 2007, pp. 77–89.

[21] J. L. Fleiss, "Measuring nominal scale agreement among many rater," Psychological Bulletin, vol. 76, 1971, pp. 378–382.

Information Hiding in Real Object Images Captured with Video Camera Using Brightness-Modulated Light

Kohei Ohshita, Hiroshi Unno, and Kazutake Uehira
Department of Information Network and Communication
Kanagawa Institute of Technology
Atsugi, Japan
e-mail: hssangaku@yahoo.co.jp, unno@nw.kanagawa-it.ac.jp, uehira@nw.kanagawa-it.ac.jp

***Abstract*— We propose a technology that can invisibly embed information into the image of a real object captured with a video camera. It uses illumination light that illuminates the real object. The illumination light invisibly contains information. As the illumination contains information, the image of the object illuminated by such light also contains information. Information in the light is produced by modulating luminance according to the embedded pattern at half-frame frequency. Frame images over a certain period are added up after the sign of the even- or odd-numbered frames is changed. Changes in brightness by modulation in each frame are accumulated over the frames while the object image is removed because the even and odd frames are opposite in sign. This makes it possible to read out the embedded patterns. This paper demonstrates the feasibility of information hiding from our experimental results.**

Keywords-information hiding; digital watermarking; brightness-modulated light.

I. Introduction

Information hiding or data embedding technologies have recently been studied and used in various applications. Although conventional information hiding technologies hide information into digital data, we have studied a technique that does not hide information in digital data but conceals it on the surface of a real object [1]–[3]. The main purpose of this technique is to protect the portrait rights of real objects that are highly valued as portraits such as those in celebrated paintings in museums. The technique we propose uses illumination light for the real objects. This light invisibly contains certain information. Since the illumination contains information, the captured images of real objects illuminated by such light also contain information as watermarks. Although we studied this technique for still images, the need for these kinds of techniques for moving images has recently been increasing because taking moving images at any time and from anywhere has recently become easier for everyone than ever before because of the widespread use of smart phones that have video cameras built into them.

The main purpose of our study was to develop the same technique for moving pictures captured with video cameras as that for still images. We propose a technique using temporally and spatially illumination-modulated light for moving images. We demonstrated the feasibility of this technique experimentally in this study by using actual moving images captured with a video camera that solved problems that arise in practical use, such as asynchronicity between video cameras and projectors.

II. Hiding and Reading Out Information

Figure 1 illustrates the basic configuration for the technique we present in this paper. It uses a kind of projector as a light source to illuminate the real object by projecting light whose luminance is temporally and spatially modulated. The embedded pattern is binary image and it is embedded in the projected light by increasing (decreasing) luminance in the patterned area by slight degree dB from the averaged value for odd- (even-) numbered frames. The amplitude of modulation, dB, is too small for human vision to perceive. Since the luminance of the light projected onto the object is modulated, the brightness of the captured image of the object is also modulated, although it is invisible. Because dB is so small, it is difficult to read out the embedded pattern from a single frame image. Therefore, frame images over certain periods are added up after the sign of the brightness in even- or odd-numbered frames is changed. Because the sign of every other frame is changed, the changes in brightness by modulation in each frame are accumulated over the frames. However, the background image is removed because the brightness of even and odd

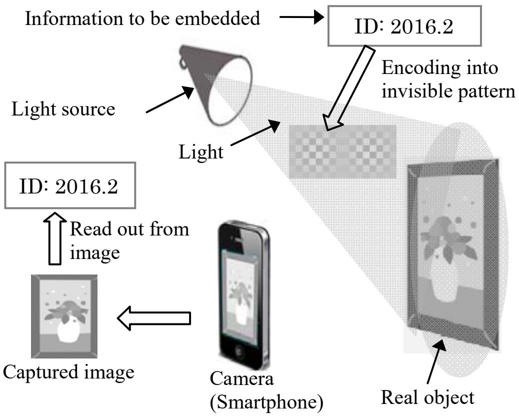


Figure 1. Basic configuration of technology.

frames cancels one another out. As a result, the embedded patterns become visible and it is then possible to read them out.

This technique presupposes that the captured image and projected image are synchronized. However, they are not actually synchronized, especially in phase. If the phase is shifted near 90 degrees in the worst case, the brightness of the modulated pattern is cancelled and cannot be added up since a frame of the image is captured across two frames of the projected moving image. To solve this problem, we propose a method where luminance is modulated at half frequency of the frame frequency of the projector that produces two groups of added up images; the first is produced by only adding up odd-numbered frames and the second is produced by only adding up even-numbered frames. The frames of captured image do not go across the two frames of projected images whose signs of dB are opposite by doing this, at least for one of these two frames. Therefore, it becomes possible to read out the embedded pattern by choosing the one with the largest contrast.

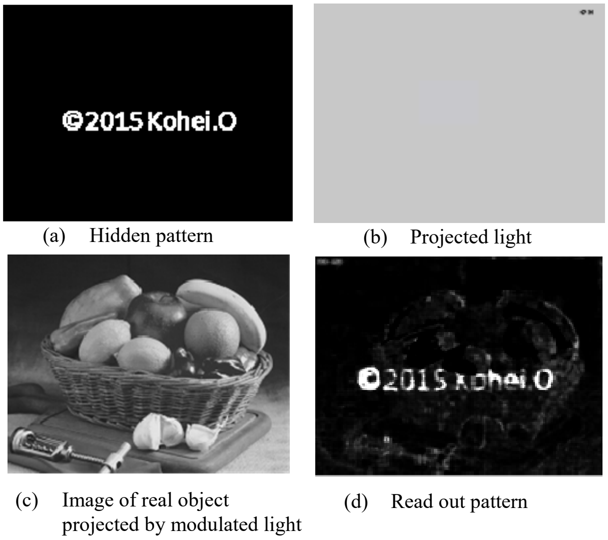
III. EXPERIMENTS

We conducted experiment to demonstrate the effectiveness of the proposed technique. We only embedded information in blue component images to enhance invisibility. We embedded simple pattern, characters, and quick response (QR) code. We used A1 size printed color images as real objects. We used a projector that had 1280 x 1024 pixels to project light that contained the invisible pattern. The brightness of the light was 200 except for the pattern embedded region where brightness was modulated with amplitude of dB, which was changed up to 20 as an experimental parameter. These figures indicate the gray scale value of which maximum is 255. The image of the object projected the light modulated with invisible pattern was captured with video camera that had 1280 x 720 pixel as a color video image.

We also evaluated the invisibility of the embedded information through subjective tests when it was in the light and captured images.

IV. RESULTS

The results from the experiments revealed that invisible patterns could be read out for dB of over four and when the number of frames (NF) added up to over 20. Figure 2(a) shows an example of the pattern to be hidden, Figure 2(b) has the brightness distribution of the projected light modulated by the pattern, Figure 2(c) has the image of the real object projected by the modulated light, and Figure 2(d) has the read out pattern. The conditions in this example were that dB was 20, NF was 30, and the size of the embedded characters in the original image was 20 pt. It can be seen in Figure 2 that the hidden pattern of the characters cannot be seen on the real object; however, it can be seen by summing up the frame images, although noise is seen in the background.



(a) Hidden pattern (b) Projected light

(c) Image of real object projected by modulated light (d) Read out pattern

Figure 2. Example of hidden pattern and read out pattern.

We found with regard to invisibility that for dB of under 10, the embedded pattern could not be seen either in the light or captured images and for dB of 20, we saw the flicker in the embedded pattern area although pattern could not be seen. These results indicated that there were conditions where both the invisibility and readability of the patterns were simultaneously satisfied although we have to remove the flicker in future.

V. CONCLUSION

We proposed a new technique for information hiding in real object images captured with a video camera using brightness-modulated light as illumination for the real object. Although the changes in brightness with this technique by modulation in each frame are too small to be visible, they become visible when they are accumulated over the frames.

We conducted experiments and the results from these revealed that invisible hidden patterns could be read out, as we had expected. We intend to remove the noise in background and flicker in the embedded pattern region in future work.

REFERENCES

[1] K. Uehira and M. Suzuki, "Digital Watermarking Technique Using Brightness-Modulated Light," Proc. of IEEE 2008 International Conference on Multimedia and Expo, Jun. 2008, pp. 257–260.

[2] Y. Ishikawa, K. Uehira, and K. Yanaka, "Practical evaluation of illumination watermarking technique using orthogonal transforms," IEEE/OSA J. Display Technology, vol. 6, No. 9, pp. 351–358, 2010.

[3] M. Komori and K. Uehira, "Optical watermarking technology for protecting portrait rights of three-dimensional shaped real objects using one-dimensional high-frequency patterns," Journal of Electric Imaging, Vol. 22, No. 3, pp. 033004-1–033004-7, 2013.

A Robust Image Watermark Algorithm for AD/DA Conversion

DongHwan Shin[1], KyungJun Park, TaeYoon Lee, ChangWon Kim and JongUk Choi
Content Solution Business Division
MarkAny Inc.
Seoul, Korea
e-mail:{dhshin[1], kjpark, tylee, permedia, juchoi}@markany.com

***Abstract*— This paper is proposing a new watermarking approach which is robust to the AD / DA (from Analog to Digital, or from Digital to Analog) conversion attack, and especially strong against print-scan attack. The new algorithm is designed to be robust for the RST (Rotation, Scaling and Translation) attack and AD/DA attack in order to be used in print-scan applications. The proposed algorithm makes the 12 circular digital signal templates in the frequency domain and shifts them circularly according to the watermark information to be embedded. In order to extract watermark, we use a method to calculate the correlation between the extracted patterns and reference patterns. As a result, our new watermark approach was robust against color print & scan attack on a paper, with 600 dpi color printer, with watermark detection rate of 83.5% in a normal condition room, and 81% in a normal outdoor condition while detection rate was 45% on a monitor display & scan attack case. The difference of detection rates was not large between indoor and outdoor environments.**

Keywords- Image; Watermark; RST; AD/DA; Circular shift.

I. INTRODUCTION

Image watermark algorithm has been used to protect copyrights of digital image content. This digital content can be distributed with embedding watermarks in it and the watermark information can be extracted on the receiver side to recognize the first downloader information or to know the copyright information of the digital image content [1][2][3][7][9]. In addition to this, two informative use cases, the watermark algorithm is recently used to inform subsidiary information of digital image contents.

Image watermark techniques are classified as spatial domain based methods and frequency domain based methods, according to the region of watermark embedding. In the spatial domain based method, the image watermark is embedded by using pixel information of the digital image. The advantage of this method is speed of embedding. It can be relatively fast because a watermark signal is embedded in a spatial domain directly. Because watermark extraction can happen from spatial domain directly, the extracting speed is also fast. Because of those advantages, many of image watermark techniques are using the spatial domain based method. However, this method is known for its weakness in attack cases, such as compression (encoding) attacks. In addition, this technique has a characteristic that it should know where the watermark information starts in a digital image to extract the watermark information. For example, there is a spread spectrum method, which is one of the famous methods in the spatial domain based approach, and one of its disadvantages is that sync signal should be used to know the exact starting point of the embedded watermark [4]. M. Kutter tried to solve this problem with a new method which adds an additional watermark pattern, beside a message watermark, to detect this sync signal against RST attack [5][6]. The disadvantage of this method was that the performance of recover logic against RST attack affects a lot the overall performance of the image watermark algorithm.

The other image watermark technique is when the image watermark is embedded in the frequency domain. There have been several different frequency transformation ways for this approach, such as DFT (Discrete Fourier Transform), DCT (Discrete Cosine Transform), and DWT (Discrete Wavelet Transform). In the frequency domain based method, before embedding the watermark, the pixel information of the image goes through the frequency domain transformation stages. After that, the watermark information is embedded on the image by using the frequency coefficient information which is the output of the transformation. The advantage of this method is that specific properties of various frequency transformation methods can be used. Furthermore, in lossy compression, such as JPEG or MPEG, various kinds of frequency transformation methods are basically used already to remove information redundancy for improving the efficiency of the compression. Because of this, the watermark embedding method in frequency domain has characteristics that make the robust watermarking approach a relatively easily technique against compression attacks.

In this paper, the proposed watermark method embeds the watermark signal on the frequency domain. DFT is used for the frequency transformation method. This method has a feature which stores the watermark information distributed over several pixels in the spatial domain, without making a huge distortion in specific pixels even if the watermark is embedded strongly in the frequency domain. In addition, by designing the watermark embedding pattern to be in regardless of the starting point of digital image, this approach could prove to be strong against cropping attacks. In this paper, the robustness of this method is proved by extracting the image watermark through taking pictures with digital cameras of mobile smartphones.

The rest of the paper is structured as follows. In Section 2, the circular shift watermark algorithm is explained. We give details on both the watermark embedding method and also the watermark extraction method. In Section 3, several tests are conducted in order to verify the effectiveness of the proposed algorithm. We conclude the paper in Section 4.

II. CIRCULAR SHIFT WATERMARK

In this paper, to embed the watermark, the method sets the standard pattern in frequency domain plane by using template and shifts circularly the pattern according to watermark information.

Figure 1 shows the example of circular shift with angle (θ) of a specific pattern. The moving phase with angle (θ) for minimum basic unit, 1, is different according to the watermark payload amount for one specific pattern. Equation (1) shows the level value to embed x bits of information in the pattern. When the watermark information is increased by 1, the phase angle moves by Δ_{θ}, as shown in equation (2).

$$level = 2^{x} \tag{1}$$

$$\Delta_{\theta} = \pi / level \tag{2}$$

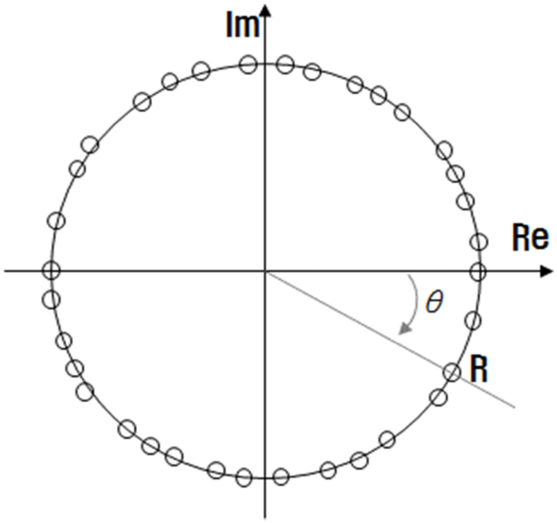


Figure 1. Watermark embedding method using Circular Shift

Figure 2 shows an example of embedded watermarks by producing multiple reference patterns.

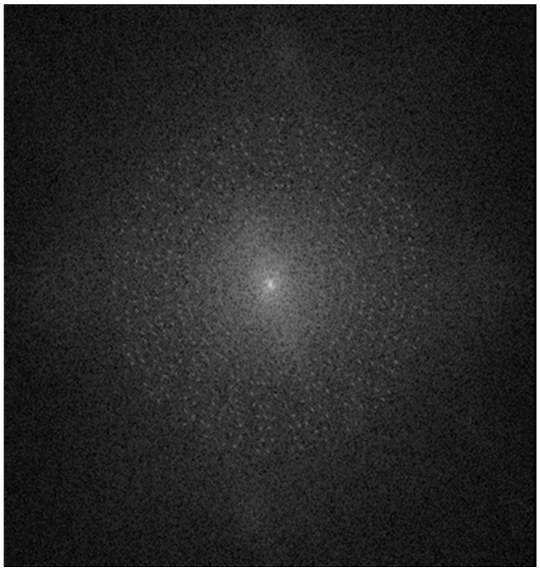


Figure 2. Example of embedding watermark in frequency plane

A. Embedding Watermark

Figure 3 shows the order of embedding the watermark.

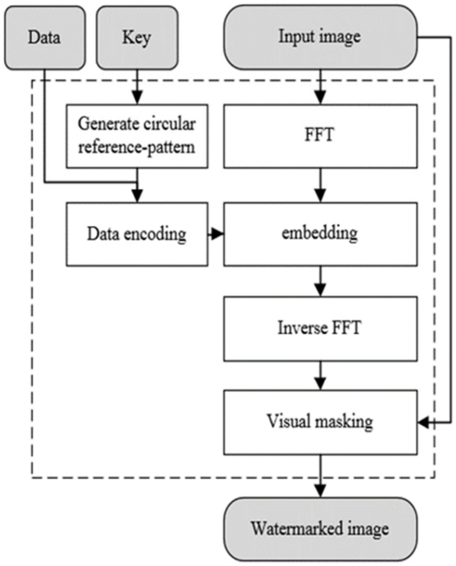


Figure 3. Order of watermark embedding

1) First, when the watermark embedding algorithm is started, a watermark message data and secret key are received as inputs, and a watermark embedding pattern data is generated based on the inputs.

2) In 1), based on the produced pattern, the watermark pattern is produced in order to be embedded in the original digital images with watermark embedding data.

3) The input digital image is transformed to frequency domain data by FFT (Fast Fourier Transform).

4) On the frequency domain of 3), the watermark pattern from 2) is embedded.

5) By using Inverse FFT, the image in the frequency domain is transformed into the image in the spatial domain.

6) To ensure the invisibility of embedded watermark, the strength of watermark is adjusted by using the visual masking method.

Equation (3) shows the visual masking method.

$$I' = WI \cdot maskoffset + (1 - maskoffset) \cdot I \qquad (3)$$

Here,

WI : the watermark embedded digital image

maskoffset : [0.0 – 1.0], if the value is closer to 0, the image is closer to the original one.

I: Original image

I' : The final image after watermark embedding

B. Extracting Watermark

Figure 4 shows the order of extracting the watermark.

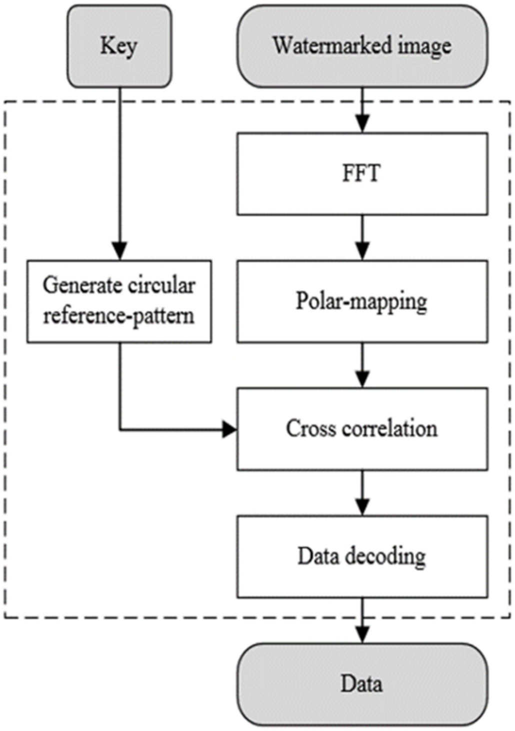


Figure 4. Order of extracting watermark

1) The watermark embedded image is transformed to frequency domain by FFT.

2) By using polar mapping, the angle component of the image is translated to horizontal axis value and the radius component of the image is transformed to vertical axis value.

3) The reference pattern is generated with the key which is used for watermark embedding.

4) The cross-correlation value is calculated between the reference pattern in 3) and the value in 2)

5) The embedded data are decoded by using the local maximum value of cross-correlation.

To extract the watermark, the correlation value between the reference pattern of each frequency and the extracted signal from each frequency band is calculated. The maximum value is selected and the phase value on the point is measured, then the watermark is extracted using the measured value. To calculate the correlation value between signals of circular form in frequency domain, a polar transformation is used to change them to a 1 dimension array of the same size.

The basic geometry of polar mapping is shown in Figure 5. Equally spaced concentric circles are drawn centered at the image center, and a number of equally spaced sectors are drawn. Pixels at the points of intersection of these circles and radial lines are plotted on a rectangular grid, and the resulting image is a polar view. In a log polar mapping, the radii of the concentric circles vary on a logarithmic scale [10][11].

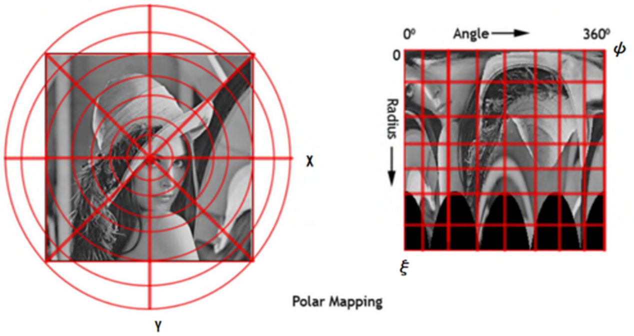


Figure 5. Comparison between rectangular coordinate system and polar coordinate system

Descartes coordinate plane:

$$z = x + yi \qquad (4)$$

Polar coordinate plane:

$$\xi = \mathrm{r}, \qquad \psi = \theta$$

$$\text{, where } r^2 = x^2 + y^2,\ \theta = \arctan(\tfrac{y}{x}) \qquad (5)$$

III. Experiment Results

Watermarks could hardly be extracted when a previous watermark algorithms, which is based on spread spectrum in spatial domain, was used for a robustness test against DA/AD conversion attack of Print-Scan method [5][12].

TABLE 1. TEST ENVIRONMENTS

Test components	Settings
Resolution	1024x1024
Printer	HP Color LaserJet4650
Camera	Samsung galaxy S4
Message bit size	66bit
PSNR	36.21dB
SSIM	0.95162

For a performance test of our algorithm, we proceeded by taking photographs in several different environments. Table 1 shows the basic test environment. The resolution of original digital image is 1024 X 1024. The printer is HP Color LaserJet 4650 for image printing. The camera is Samsung Galaxy S4 to take pictures. 66 bits of watermark payload are embedded and extracted from the taken pictures. The watermark payload does not include any error correction or error detection code.

To evaluate the visual quality of the watermark embedded image, PSNR (Peak Signal to Noise Ratio) and SSIM (Structured Similarity Index Measure) are calculated.

o PSNR – Peak Signal-to-Noise Ratio :

$$PSNR = 10 \cdot log_{10}(\frac{255^2}{MSE}) \qquad (6)$$

$$, where \quad MSE = \frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}[x(i,j) - y(i,j)]^2$$

o SSIM – Structured Similarity Index Measure

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \qquad (7)$$

$$, where$$
$$c_1 = (k_1L)^2, c_2 = (k_2L)^2$$
$$L = the\ dynamic\ range\ of\ the\ pixel\ values$$
$$k_1 = 0.01, \ \ k_2 = 0.03\ by\ default$$

For taking pictures, Samsung Galaxy S4 is used. Table 2 shows the test conditions of: picture resolution, picture object, and picture place. Two image resolutions of 1280x720 and 1920x1080 are used. The output resolution for rendering in monitor is 1920x1080. In the case of printing papers, a color printer with 600 dpi is used. In capturing from a monitor, only indoor light condition is target for test. However, in capturing from the printed papers, both indoor light and outdoor light are used for the test.

TABLE 2. CAPTURE ENVIRONMENTS

Resolution	Object	Place (experiment environment)
1280x720	Monitor (resolution: 1920x1080)	Indoor
2M pixels (1920 x1080 included)	Prints(color/resolution : 600 dpi)	Indoor/outdoor

Figure 6 shows the digital images which are used for tests. The test digital images are globe image, landscape, and video color bar image.

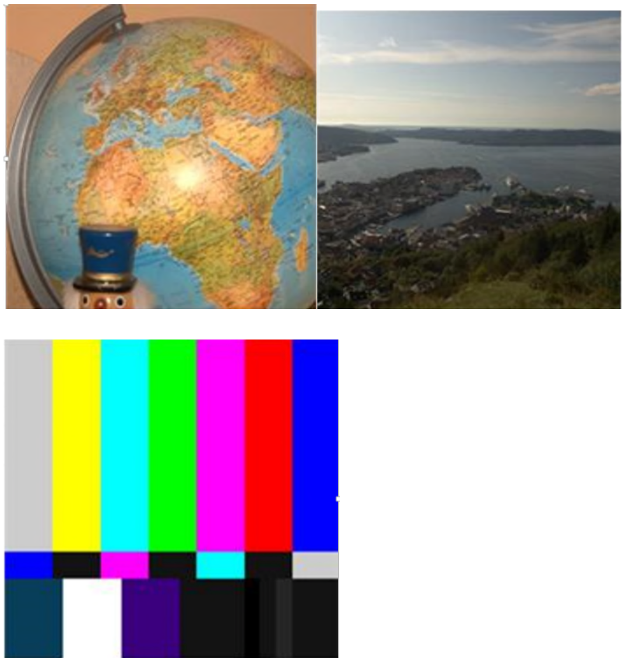


Figure 6. Digital images used for tests

Figure 7 shows the example of extracting the watermark process: (a) is the result of 2 dimension FFT of the input digital image. To embed 66 bits, a total of 12 circular patterns are used with 11 patterns for message and 1 pattern for reference (b) shows 3 dimensional picture which renders the result of the cross correlation between the reference pattern and entire image values. It shows that the locations of maximum peak for phase are different according to the watermark information (c) shows the changes of maximum correlation values according to radius changes (d) shows the correlation value with the reference pattern.

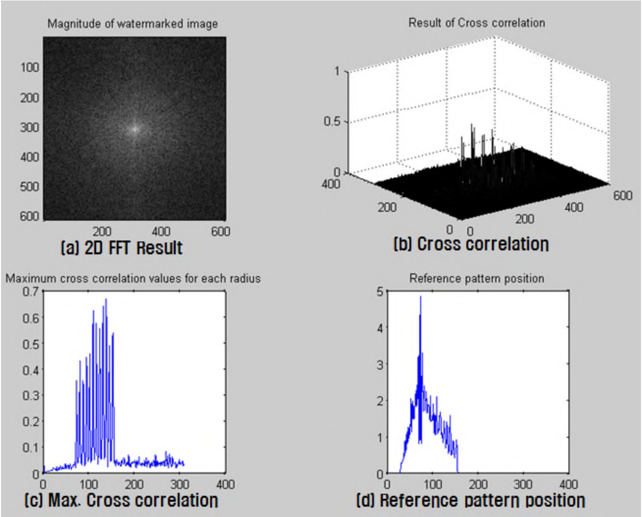
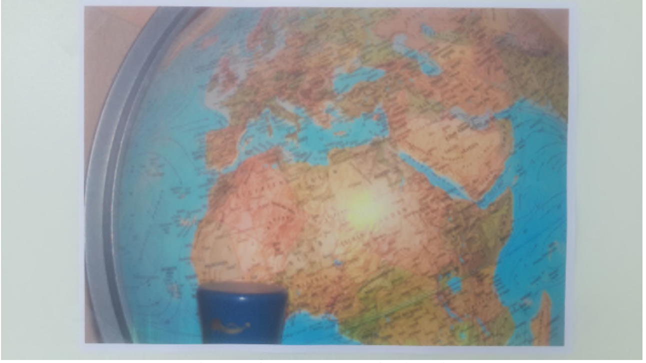


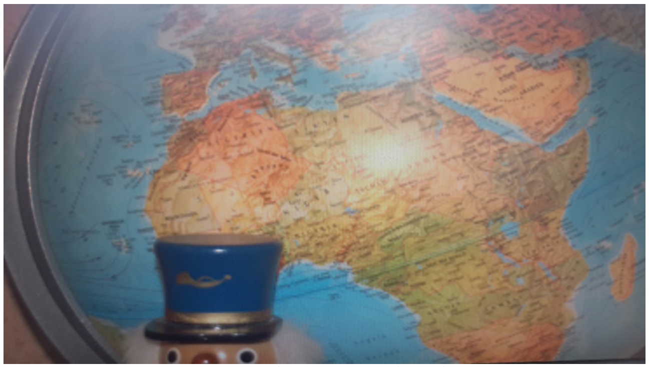
Figure 7. Example of extracting watermark

Figure 8 shows the captured image for watermark detection tests (a) is the camera captured image of the printed paper with watermark embedded in indoor environment (b) is the camera captured image of the printed paper with watermark embedded in outdoor environment (c) is the camera captured image of monitor screen with watermark embedded in indoor environment.

(a) Indoor capture of printed paper



(b) Outdoor capture of printed paper



(c) Indoor capture of monitor

Figure 8. Example of the captured images in test

Table 3 shows the test results to compare the performance of the proposed and spread spectrum method. The spread spectrum method was implemented by Kutter's method [5][6]. Success cases indicate no-error extraction cases with 66 bit data size.

The performance of the algorithm is tested with the captured images of monitor and captured images of the printed images.

When the watermarked digital image is displayed through indoor monitor, the results of extraction rate of same resolution capture is 32% and with 720p resolution is 58%.

In the case of 600 dpi color printed on paper, extraction rate of 1920x1080 resolution is 83% and with 720p is 84% in indoor cases. In the outdoor case, those were 81% and 79% for 1080p and 720p cases respectively.

TABLE 3. EXPERIMENTAL RESULTS

Test	**Environment**	**Proposed Result(Success cases/entire tries, extract rate)**	**Spread spectrum Method**
Indoor print	Color printed in 600 dpi, Captured resolution : 1920x1080	83/100 (83%)	0/100 (0%)
Indoor print	Color printed in 600 dpi, Captured resolution : 1280x720	84/100 (84%)	0/100 (0%)
Outdoor print	Color printed in 600 dpi, Captured resolution : 1920x1080	83/100 (83%)	0/100 (0%)
Outdoor print	Color printed in 600 dpi, Captured resolution : 1280x720	79/100 (79%)	0/100 (0%)
Indoor monitor	Monitor resolution : 1920x1080, Captured resolution : 1920x1080	16/50 (32%)	3/100 (3%)
Indoor	Monitor resolution:	29/50	2/100

monitor	1920x1080, Captured resolution : 1280x720	(58%)	(2%)

IV. Conclusion

In this paper, a strong watermark method against AD/DA transformation is proposed. In previous watermark algorithm approach in spatial domain, there is a disadvantage that the algorithm is weak at prints-scan attacks, which is one very strong AD/DA attack. Our algorithm is proposed to overcome this disadvantage and can be extracted regardless of RST attack. This is implemented through a method that generates a reference template watermark pattern in the frequency domain, and extracts watermark using maximum peak location by calculating the correlation between reference pattern and the extracted pattern from the watermarked image.

For evaluating the performance of the algorithm, tests were performed by capturing monitor output and printed images and extracting of watermarks. In the case of monitor output, our algorithm achieved 45% extraction rate on average. In the case of 600dpi color printer, the extraction rate is 83.5% in indoor case, and 81% in outdoor case. Because those results are without error correction mechanisms, we can expect higher extraction rates if we apply it in the future. The performance of proposed algorithm is affected by the capture angle. Further research is required in this area.

Acknowledgment

This research project was supported by Government Fund from Korea Copyright Commission in 2015. (Development of high performative watermarking embedding and detecting system for copyright protection of UHD real-time broadcasting content.).

References

[1] J. Nah, J. Kim, and J. Kim, “Video Forensic Marking Algorithm Using Peak Position Modulation,” Journal of Applied Mathematics & Information Sciences (AMIS), Vol. 6, No. 3S, pp.2391-2396, 2013.

[2] D. Li, and J. Kim, “Secure Image Forensic Marking Algorithm using 2D Barcode and Off-axis Hologram in DWT-DFRNT Domain, ” Applied Mathematics and Information Sciences, Vol. 6, 2S, pp. 513-520, 2012.

[3] J. Nah, J. Kim, and J. Kim, “Image Watermarking for Identification Forgery Prevention,” Journal of the Korea Contents Association, Vol. 11, No. 12, pp.552-559, 2011.

[4] S. P. Maity, P. K. Nandi, and T. S. Das, “Robustness improvement in spread spectrum watermarking using M-ary modulation,” Proc. 11th NationalConferenceonCommunication, NCC2005, pp.569-573, 2005.

[5] M. Kutter, “Performance improvement of spread spectrum based image watermarking schemes through M-ary modulation”, Lecture Notes in Computer Science, 1728, pp.238~250, 2000.

[6] M. Kutter, “Watermarking resistant to translation, rotation and scaling,” in Proc. SPIE, Int. Symp. Multimedia Systems and Applications, vol. 3528, pp. 423–431, 1998.

[7] C. Lin, C Chang, and Y. Chen, “A Novel SVD-based Watermarking Scheme for Protecting Rightful Ownership of Digital Images,” Journal of Information Hiding and Multimedia Signal Processing, Vol.5, No.2, pp.124-143, April 2014

[8] Makhloghi, M., Tab. F. A., and Danyali, H. “A new robust blind DWT-SVD based digital image watermarking”, ICEE 2011, Vol.1, pp.1-5, 2011.

[9] M. Wu, and B. Liu, “Data hiding in image and video: Part I—Fundamental issues and solutions,” IEEE Trans. Image Process., vol. 12, no. 6, pp. 685−695, Jun. 2003.

[10] P. Pu, X. Guo, and L. Lei., “Application of Image Interpolation in Log-Polar Transformation. Computer Engineering,” 34, 5 (2008), pp.198-199.

[11] B. Yu, L. Guo, and T. Zhao, “Gray projection image stabilizing algorithm based on log-polar image transform. Computer Applications,” 28, 12 (2008), pp.3126-3128.

[12] Y. Xin, and M. Pawlak, “M-ary Phase Modulation for Digital Watermarking” Int. J. Math. Comput. Sci. 2008, Vol. 18, No. 1, pp.93-104.

Technique for Protecting Copyrights of Digital Data for 3-D Printing, and Its Application to Low Infill Density Objects

Masahiro Suzuki, Piyarat Silapasuphakornwong,
Hideyuki Torii, Hiroshi Unno, Kazutake Uehira
Kanagawa Institute of Technology
Atsugi, Japan
E-mail: msuzuki@ctr.kanagawa-it.ac.jp,
silpiyarat@gmail.com, torii@nw.kanagawa-it.ac.jp,
unno@nw.kanagawa-it.ac.jp, uehira@nw.kanagawa-it.ac.jp

Youichi Takashima
NTT Service Evolution Laboratories
Nippon Telegraph and Telephone Corporation
Yokosuka, Japan
E-mail: takashima.youichi@lab.ntt.co.jp

***Abstract*— We evaluated our previously proposed technique to protect copyrights of digital data for 3-D printing. The technique embeds copyright information into not only digital data but also fabricated objects and enables the information to be reads to reveal possible copyright violations. In this study, copyright information was embedded into low infill density objects by constructing their inside with high infill density areas, and thermography was used for readout. An experiment was conducted to examine whether high infill density areas can be recognized from thermal images. The results indicated that they can be. We demonstrated that our technique is applicable to low infill density objects.**

Keywords-digital fabrication; 3-D printing; 3-D printer; copyright; digital watermarking.

I. INTRODUCTION

Three-dimensional (3-D) printers have become popular with consumers. People who have 3-D printers can easily fabricate products by simply obtaining the digital data for 3-D printing. Hence, many people believe that 3-D printers will change the ways in which manufacturing and physical distribution will be carried out in the near future [1] [2].

However, such benefits of 3-D printers mean that anyone can easily manufacture bootleg products if he or she abuses digital data for 3-D printing. Such copyright violations will obviously cause serious economic damage. Thus, techniques to protect the copyrights of digital data for 3-D printing are essential for the healthy development of markets for 3-D printers.

Although techniques to prevent illegal copying or illegal printing are of course important to protect the copyrights of digital data for 3-D printing [3]–[6], techniques to divulge such violations are additionally crucial in cases in which these violations occur. Digital watermarking is common to all kinds of digital data to disclose copyright violations; however, conventional techniques can only embed watermarks into digital data but not into actual objects fabricated with 3-D printers. Copyright information needs to be embedded into fabricated objects so that the information can be read to reveal any violations. For example, when a company that is not allowed to use digital data held by the copyright holder is selling illegally fabricated objects, the copyright holder can expose the company by detecting the copyright information embedded into the sold objects and can assert his or her just rights. Thus, techniques to reveal copyright violations, i.e., techniques to embed copyright information into fabricated objects and have the information be readable, are essential to protect the copyrights of digital data for 3-D printing.

We have previously proposed a technique to embed copyright information into fabricated objects and to have that information be readable and demonstrated its feasibility [7] [8]. With this technique, the inside of objects is constructed with small areas, which have physical characteristics that are different from the areas surrounding them, to embed the information. These small areas are detected using nondestructive inspections, such as X-ray photography or thermography, to reveal the embedded information. Our previous studies demonstrated that information embedded with small cavities inside objects can be read. Thus, information can feasibly be embedded into objects and be read.

Although we have evaluated our technique using objects fabricated with high infill density, i.e., 100% [7] [8], objects with low infill density are often fabricated for practical use. For example, when users want to save time or materials, they fabricate objects with low infill density. In our previous studies, we embedded copyright information into high infill density objects with small cavities inside because they did not have any other empty spaces; however, low infill density objects have many empty spaces inside, and embedding the information into them is difficult. Thus, methods applicable to low infill density objects are required for practical use with our technique.

This paper presents a method of applying our technique to low infill density objects. In Section 2, we first describe the basic concept of the technique and the principle of the methods previously applied to high infill density objects then describe the principle of our method for low infill density objects. In Section 3, we describe the methodology, explain the results, and provide discussion of an experiment to evaluate the feasibility of our method. In Section 4, we conclude the paper.

II. Our Technique and Its Application to Low Infill Density Objects

Figure 1 is an illustration showing the basic concept of our technique. Digital data for 3-D printing are created with 3-D modeling software, such as 3-D CAD, and copyright information is embedded into the data. When actual objects are fabricated using the data, the information embedded into the data is also embedded into the objects. This information is made readable using nondestructive inspection, such as X-ray photography or thermography, which enable the detection of copyright violations, such as making bootleg products. Thus, the copyrights of digital data for 3-D printing are protected by embedding copyright information into not only digital data but also fabricated objects for readability.

Figure 2 is an illustration showing the principle of our technique. In our previous studies, copyright information, which was expressed using ASCII, was encoded with small cavities inside high infill density objects [7] [8]. For example, the existence and nonexistence of cavities at certain positions were represented as “0” and “1,” respectively (see Figure 2-A). The objects, which had cavities inside them, were heated using electric equipment, such as halogen lamps, and the surface temperature of the objects was measured using thermography (see Figure 2-B). The surface temperature at positions where the cavities existed became higher than that at other positions because the cavities blocked thermal flow (see Figure 2-B). Previous studies demonstrated that the existence or nonexistence of cavities can be detected from thermal images and that the embedded information can be decoded from the detected cavities [7] [8]. Thus, thermography can feasibly be used to read copyright information embedded into high infill density objects by constructing their insides with small cavities.

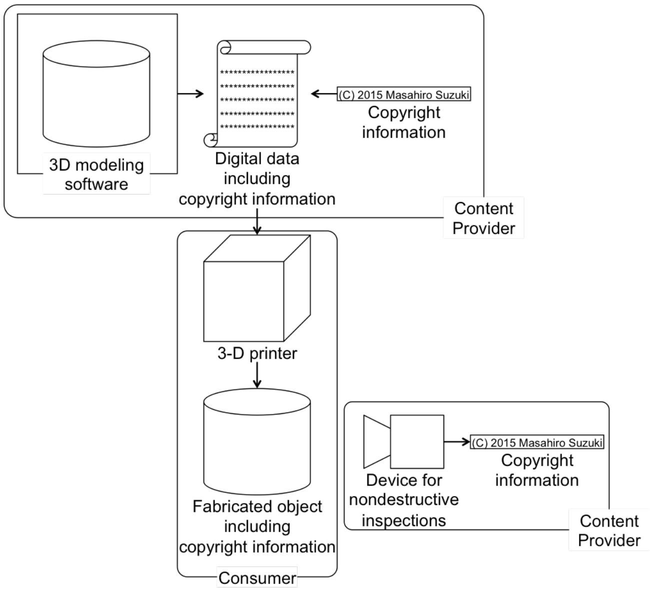


Figure 1. Illustration showing basic concept of our technique.

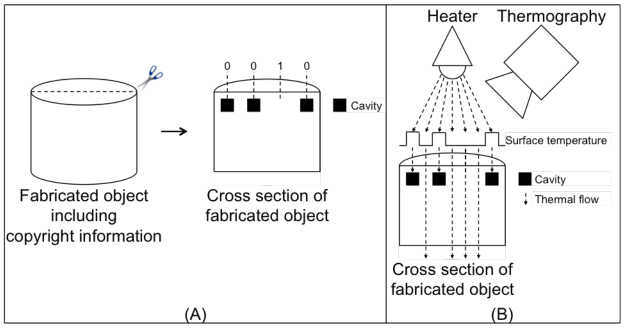

Fabricated object including copyright information
0 0 1 0
Cavity
Cross section of fabricated object
Heater
Thermography
Surface temperature
Cavity
Thermal flow
Cross section of fabricated object
(A)
(B)

Figure 2. Illustration showing principle of our technique.

However, because low infill density objects have many empty spaces inside, the aforementioned method cannot be applied. For example, the low infill density objects shown in Figure 3 have a honeycomb structure inside, and each space inside the honeycomb is empty. If we embed copyright information into such objects with small cavities inside and heat them, the surface temperature in the cavities would not differ from other positions because not only the cavities but also the empty spaces would block thermal flow. Thus, other methods are required for applying our technique to low infill density objects.

We present a method for embedding copyright information into low infill density objects and having that information be readable. Figure 4 is an illustration showing the principle of this method. Copyright information expressed using ASCII is encoded with high infill density areas, i.e., pillars, inside low infill density objects. The

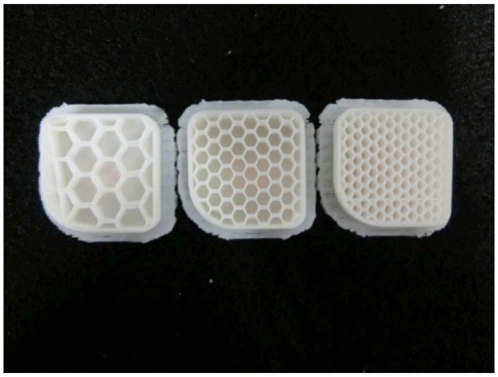


Figure 3. Photographs of inside low infill density objects.

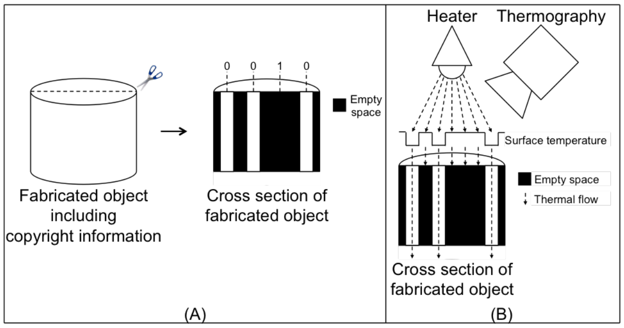

Fabricated object including copyright information
0 0 1 0
Empty space
Cross section of fabricated object
Heater
Thermography
Surface temperature
Empty space
Thermal flow
Cross section of fabricated object
(A)
(B)

Figure 4. Illustration showing principle of method for applying our technique to low infill density objects.

existence and nonexistence of the pillars is represented as “0” and “1,” respectively. The objects are heated, and the surface temperature is measured using thermography. The surface temperature at the positions where the pillars exist becomes higher than that of the other positions because the pillars do not block thermal flow, even though empty spaces do. The existence or nonexistence of pillars is detected from thermal images, and the embedded information is decoded from the detected pillars. Thus, copyright information is embedded into low infill density objects by constructing their inside with high infill density areas and the information is made readable using thermography.

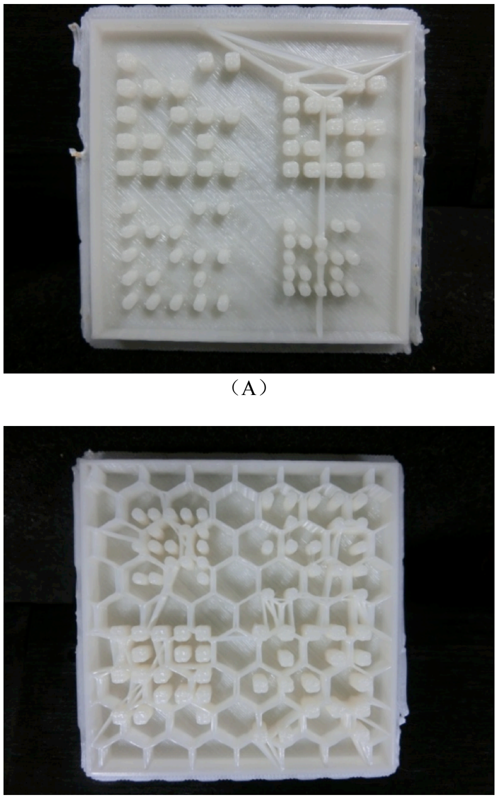
III. Experiment

A. Methodology

An experiment was conducted to examine whether high infill density areas, i.e., pillars, inside low infill density objects can be recognized from thermal images. The recognition rate was calculated using the following equation:

$$R_{\text{recognition}}(\%) = \frac{N_{\text{detected}}}{N_{\text{embedded}}} \times 100\,, \qquad (1)$$

where the $R_{\text{recognition}}$ is the recognition rate, the N_{embedded} is the number of existing or non-existing pillars used to embed copyright information, and N_{detected} is the number of existing or non-existing pillars correctly detected. The N_{detected} was determined by analyzing thermal images of two test samples, which were fabricated from black polylactic acid filament using a fused deposition modeling 3-D printer, i.e., a MakerBot Replicator. The first test sample had 1% infill density, and the second test sample had 10% infill density (see Figure 5). Each test sample had four arrays of pillars, and each array corresponded to each of four conditions: two conditions of pillar size (1 and 2 mm) × two conditions of space between the pillars (1 and 2 mm) (see Figure 6). In an analysis of the thermal images, a still image was extracted from a video image for each frame, and each still image was binarized using adaptive thresholding [9]. A logical disjunction among binarized images was calculated for each position where copyright information was embedded, and N_{detected} was defined as the number of positions where the logical disjunction accorded with the existing or non-existing pillar. The thermal images were taken using Testo 875. Two halogen lamps, positioned at 16 cm and 60 degrees left and right, were used to heat the test samples. In each trial of taking the thermal images, the test samples were first heated, then the lamps were turned off. The test samples were then cooled to their initial temperature. The thermal images from turning the lamps off were used for the analysis. Thus, the recognition of high infill density areas inside low infill density objects from thermal images was examined by calculating the recognition rate.


(A)

(B)

Figure 5. Photographs of inside test samples. (A) Inside of test sample whose infill density was 1%. (B) Inside of test sample whose infill density was 10%. Note that test samples actually used in experiment were black.

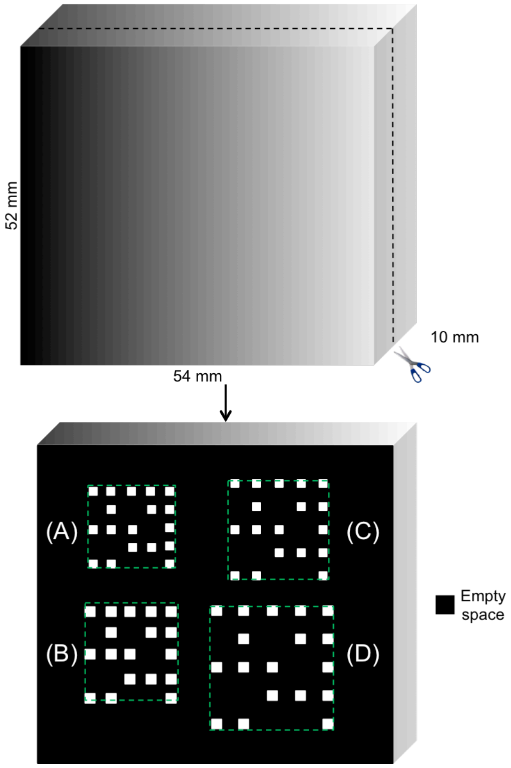
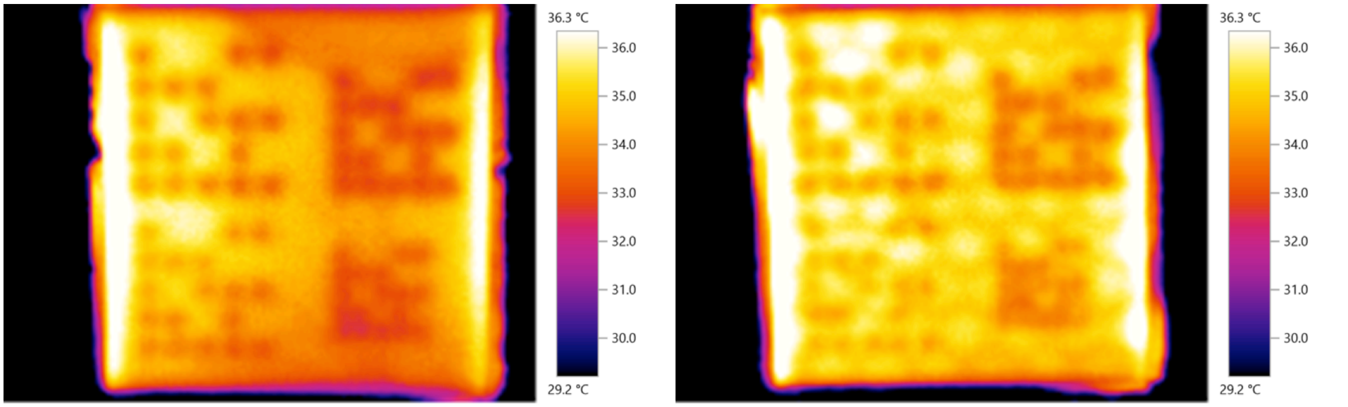


Figure 6. Schematic of test samples. (A) Array of pillars under 1-mm size × 1-mm space condition. (B) Array of pillars under 2-mm size × 1-mm space condition. (C) Array of pillars under 1-mm size × 2-mm space condition. (D) Array of pillars under 2-mm size × 2-mm space condition.



(A)

(B)

Figure 7. Results of thermal images of test samples. (A) Thermal image of test sample whose infill density was 1%. (B) Thermal image of test sample whose infill density was 10%.

TABLE 1. RESULTS OF THE RECOGNITION RATE (%)

Infill density							
1%				***10%***			
Size				**Size**			
1 mm		***2 mm***		***1 mm***		***2 mm***	
Space		**Space**		**Space**		**Space**	
1 mm	***2 mm***	***1 mm***	***2 mm***	***1 mm***	***2 mm***	***1 mm***	***2 mm***
100	100	100	100	88	100	96	100

B. Results and Discussion

Figure 7 shows the thermal images, and Table 1 summarizes the recognition rate. All four arrays were completely recognized under the 1% condition, and the two arrays with the 2-mm space were completely recognized under the 10% condition. These results suggest that copyright information can be embedded into low infill density objects by constructing their inside with high infill density areas and be readable using thermography. Thus, our technique can feasibly be applied to low infill density objects.

IV. CONCLUSION

We evaluated our technique to protect copyright information of digital data for 3-D printing. This technique embeds copyright information into low infill density objects by using our method of constructing their inside with high infill density areas and makes the information readable using thermography. We conducted an experiment to examine whether high infill density areas inside low infill density objects can be recognized from thermal images. The results indicated that they can be. Thus, we demonstrated that our technique is feasible.

ACKNOWLEDGMENT

This study was supported by JSPS KAKENHI Grant Number 15H02707.

REFERENCES

[1] B. Berman, “3-D printing: The new industrial revolution,” Business horizons, vol. 55, no. 2, March–April 2012, pp. 155–162.

[2] B. Garrett, “3D printing: new economic paradigms and strategic shifts,” Global Policy, vol. 5, no. 1, February 2014, pp. 70–75.

[3] P. R. Alface and B. Macq, “From 3D mesh data hiding to 3D shape blind and robust watermarking: a survey,” Transactions on Data Hiding and Multimedia Security II, 2007, pp. 91–115.

[4] T. Modegi, “A proposal of 3D-printing regulation technique for fabricating hazardous devices or copyright violating objects using by polygon matching algorithm,” IEICE Technical Report, vol. 114, no. 33, May 2014, pp. 23–28.

[5] T. Modegi, “Proposal for 3D-printing regulation technique in fabricating dangerous devices,” Bulletin of the Technical Association of Graphic Arts of Japan, vo. 51, no. 4, 2014, pp. 246–255.

[6] K. Wang, G. Lavoue, F. Denis, and A. Baskurt, “A comprehensive survey on three-dimensional mesh watermarking,” IEEE Transactions on Multimedia, vol. 10, no. 8, December 2008, pp. 1513–1527.

[7] A. Okada et al., “Non-destructively reading out information embedded inside real objects by using far-infrared light,” Proc. SPIE 9599, Applications of Digital Image Processing XXXVIII, September 2015, pp. 95992V-1–95992V-7.

[8] M. Suzuki, P. Silapasuphakornwong, K. Uehira, H. Unno, and Y. Takashima, “Copyright protection for 3D printing by embedding information inside real fabricated objects,” International Conference on Computer Vision Theory and Applications, 2015, pp. 180–185.

[9] T. R. Singh, S. Roy, O. I. Singh, T. Sinam, and K. M. Singh, “A new local adaptive thresholding technique in binarization,” International Journal of Computer Science Issues, vol. 8, no. 6, November 2011, pp. 271–277.