SIGNAL 2019

The Fourth International Conference on Advances in Signal, Image and Video Processing

ISBN: 978-1-61208-716-0

June 2 - 6, 2019

Athens, Greece

SIGNAL 2019 Editors

Wilfried Uhring, Université de Strasbourg, France

SIGNAL 2019

Forward

The Fourth International Conference on Advances in Signal, Image and Video Processing (SIGNAL 2019), held between June 02, 2019 to June 06, 2019 - Athens, Greece, continued a series of events related to signal, image and video processing.

Signal, video and image processing constitutes the basis of communications systems. With the proliferation of portable/implantable devices, embedded signal processing became widely used, despite that most of the common users are not aware of this issue, New signal, image and video processing algorithms and methods, in the context of a growing-wide range of domains (communications, medicine, finance, education, etc.) have been proposed, developed and deployed. Moreover, since the implementation platforms experience an exponential growth in terms of their performance, many signal processing techniques are reconsidered and adapted in the framework of new applications. Having these motivations in mind, the goal of this conference was to bring together researchers and industry and form a forum for fruitful discussions, networking, and ideas.

We welcomed academic, research and industry contributions. The conference had the following tracks:

- Features and models for images/signals
- Special signal, image and video processing applications/domains
- Signal processing theory and methods

We take here the opportunity to warmly thank all the members of the SIGNAL 2019 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 who dedicated much of their time and effort to contribute to SIGNAL 2019. We truly believe that, thanks to all these efforts, the final conference program consisted of top quality contributions.

We also thank the members of the SIGNAL 2019 organizing committee for their help in handling the logistics and for their work that made this professional meeting a success.

We hope that SIGNAL 2019 was a successful international forum for the exchange of ideas and results between academia and industry and to promote further progress in the areas of signal, image and video processing. We also hope that Athens, Greece provided a pleasant environment during the conference and everyone saved some time to enjoy the historic charm of the city.

SIGNAL 2019 Chairs

SIGNAL Steering Committee

Wilfried Uhring, Université de Strasbourg, France
G. Sahoo, BIT Mesra, Ranchi, India
Malka N. Halgamuge, University of Melbourne, Australia

SIGNAL 2019
Committee

SIGNAL Steering Committee
Wilfried Uhring, Université de Strasbourg, France
G. Sahoo, BIT Mesra, Ranchi, India
Malka N. Halgamuge, University of Melbourne, Australia
Laurent Fesquet, Grenoble Institute of Technology - TIMA, France
Jérôme Gilles, San Diego State University, USA
Constantin Paleologu, Polytechnic University of Bucharest, Romania
Zhongyuan Zhao, Beijing University of Posts and Telecommunications, China
Demetrios Sampson, Curtin University, Australia
Andrea Kutics, International Christian University, Japan
Pavel Loskot, Swansea University, UK

SIGNAL Industry/Research Advisory Committee
Sergey Y. Yurish, Excelera, S. L. | IFSA, Spain
Filippo Vella, National Research Council of Italy, Italy
Jai Gopal Pandey, CSIR-CEERI (Gov. of India), India
Tudor-Catalin Zorila, Toshiba Cambridge Research Laboratory, UK

SIGNAL 2019 Technical Program Committee

Waleed H. Abdulla, The University of Auckland, New Zealand
Afaq Ahmad, Sultan Qaboos University, Oman
Kiril Alexiev, Institute for Information and Communication Technologies -Bulgarian Academy of Sciences, Bulgaria
Hamada Alshaer, University of Edinburgh, UK
Cristian Anghel, Politehnica University of Bucharest, Romania / Pentalog, France
Vijayan K. Asari, University of Dayton, USA
Abdourrahmane M. Atto, University Savoie Mont Blanc, France
Nadia Baaziz, Université du Québec en Outaouais, Canada
Junaid Baber, Asian Institute of Technology, Thailand
Vesh Raj Sharma Banjade, Intel Coporation, USA
Denis Beautemps, CNRS | GIPSA-lab, France
Haithem Ben Chikha, Tunisia Polytechnic School, Tunisia
Wassim Ben Chikha, Tunisia Polytechnic School, Tunisia
Stefano Berretti, University of Florence, Italy
Silvia Biasotti, CNR - IMATI, Italy
Jacques Blanc-Talon, DGA, France
Larbi Boubchir, LIASD - University of Paris 8, France
Abdel-Ouahab Boudraa, Ecole Navale/Arts & Métiers ParisTech, France
Samia Boukir, Bordeaux INP (Bordeaux Institute of Technology), France

Copyright Information

For your reference, this is the text governing the copyright release for material published by IARIA.

The copyright release is a transfer of publication rights, which allows IARIA and its partners to drive the dissemination of the published material. This allows IARIA to give articles increased visibility via distribution, inclusion in libraries, and arrangements for submission to indexes.

I, the undersigned, declare that the article is original, and that I represent the authors of this article in the copyright release matters. If this work has been done as work-for-hire, I have obtained all necessary clearances to execute a copyright release. I hereby irrevocably transfer exclusive copyright for this material to IARIA. I give IARIA permission or reproduce the work in any media format such as, but not limited to, print, digital, or electronic. I give IARIA permission to distribute the materials without restriction to any institutions or individuals. I give IARIA permission to submit the work for inclusion in article repositories as IARIA sees fit.

I, the undersigned, declare that to the best of my knowledge, the article is does not contain libelous or otherwise unlawful contents or invading the right of privacy or infringing on a proprietary right.

Following the copyright release, any circulated version of the article must bear the copyright notice and any header and footer information that IARIA applies to the published article.

IARIA grants royalty-free permission to the authors to disseminate the work, under the above provisions, for any academic, commercial, or industrial use. IARIA grants royalty-free permission to any individuals or institutions to make the article available electronically, online, or in print.

IARIA acknowledges that rights to any algorithm, process, procedure, apparatus, or articles of manufacture remain with the authors and their employers.

I, the undersigned, understand that IARIA will not be liable, in contract, tort (including, without limitation, negligence), pre-contract or other representations (other than fraudulent misrepresentations) or otherwise in connection with the publication of my work.

Exception to the above is made for work-for-hire performed while employed by the government. In that case, copyright to the material remains with the said government. The rightful owners (authors and government entity) grant unlimited and unrestricted permission to IARIA, IARIA's contractors, and IARIA's partners to further distribute the work.

Table of Contents

Smoke Detection Using GMM and Deep Belief Network

Rabeb Kaabi
ENSIT, LR13ES03, SIME
Université de Tunis
Montfleury, Tunisia
rabebkaabi89@gmail.com

Mounir Sayadi
ENSIT, LR13ES03, SIME
Université de Tunis
Montfleury, Tunisia
mounirsayadi@yahoo.fr

Moez Bouchouicha
LIS-CNRS
Université de Toulon-Université Aix-Marseille
Toulon-Marseille, France
moez.bouchouicha@univ-tln.fr

Eric Moreau
Université de Toulon-Université Aix-Marseille
Toulon-Marseille, France
eric.moreau@univ-tln.fr

***Abstract*— The objective of this work is to develop a deep learning model for classification of smoke and no smoke regions in aerial recorded videos. For that, a deep belief network model was selected and implemented. First, frames were extracted from the provided videos. The Gaussian Mixture Model (GMM) was applied as background estimation algorithm. Then, the Deep Belief Network algorithm was applied to detect the smoke for the candidate region. Deep Belief Network was implemented and tested on different datasets. Overall, the obtained results reveal that our implemented model was able to accurately classify smoke and no smoke regions. Through the experiments with input videos obtained from various weather conditions, the proposed algorithms were useful to detect smoke in forests to minimize the damage caused by forest fires onto vegetation, animals and humans.**

Keywords- smoke detection; GMM; Deep Belief Network.

I. INTRODUCTION

Video-based smoke detection systems are composed of two types of methods. The first one relies on static features, such as the color of the smoke. The second method uses dynamic feature like movement, texture, etc.

Our proposed method of smoke detection uses dynamic features based on the Gaussian Mixture Model [1] and the Deep Belief Network (DBN) [2][3] as a classifier to reduce the false alarms and to increase the detection rate of video smoke detection systems.

Several researchers have played a significant role in the development of useful video smoke detection systems. We focused on some of them. Hu et al. [4] extract the motion feature and use the Convolutional Neural Network (CNN) [5] as classifier. Chen et al. [6] extract motion, color and energy features to classify smoke and no smoke regions using Support Vector Machine SVM [7]. Toreyin et al. [8] extract motion, color, energy and texture as features and classify the smoke and no smoke regions using decision trees.

The rest of the paper is organized as follows: the proposed technique intended for smoke detection is presented and detail in Section 2. Experimental results are shown in Section 3. Conclusion and perspectives are presented in the last section.

II. THE PROPOSED METHOD

A. Smoke Motion Detection

First, extracting candidated regions is a crucial step to know the nature of motion (ordinary or chaotic). Labeling motion regions could be done by using three methods: Optical Flow, Background Subtraction and Temporal differencing. The technique intended for background subtraction that we used in simulations is the Gaussian Mixture Model [1]. This method subtracts the background image from the current frame to find regions containing motions. In this approach, the camera is stationary.

Each pixel in the frame is defined by a mixture of K Gaussian distributions. The probability that a pixel represents the intensity is defined by:

$$P(I_t) = \sum_{i=1}^{K} w_i \gamma(I_t, \mu_i, \sigma_i^2) \quad (1)$$

where w_i is the weight.

μ_i is the mean.

σ_i^2 is the covariance for the i th distribution.

γ is a Gaussian probability density function:

$$\gamma(I_t, \mu_i, \sigma_i^2) = \frac{e^{-1/2\sigma^2 (I_t - \mu)^T (I_t - \mu)}}{(2\pi)^{1/2}\sigma} \quad (2)$$

To model the background, Bk is estimated as:

$$B_k = \arg\min(\sum_{i=}^{b} w_i > Th) \quad (3)$$

Th is the minimum fraction of the background model.

 ISBN: 978-1-61208-716-0

B. *Fisher Vector*

Fisher vector [9] represents a dimension reduction technique that can be used for classification as well. It picks a new dimension that gives maximum separation between means of projected classes and minimum variances within each projected class. The Fisher Vector is an image representation obtained by pooling local image features. It is mostly used as an image descriptor in visual classification and improves the classification performance of the representation. Our developed idea is realized as follows:

- Fitting a Gaussian Mixture Model (GMM)
- Saving and loading the fitted GMM
- Computing the Fisher Vectors based on the fitted GMM

C. *Deep Belief Network*

Deep Belief Network is a tool of machine learning that represents a stack of Restricted Boltzmann Machine [10]. After the pre-training process with the RBMs, the network acts like a multi-layer Perceptron [11] using the backpropagation [12] as tool to accomplish the training. The architecture of the used Deep Belief Network is presented in Figure 1.

Our aim is to improve the performance of smoke detection. An efficient method is presented using deep belief network. The algorithm is separated into two major phases. The first phase is to segment the candidate regions in video sequences. The Fisher criterion is applied to maximize the ratio of the separation of the two classes with respect to their dispersions. The Fisher criterion [9] is similar to Principle components analysis but it focuses on maximizing the separability among known categories. Second, the normalized candidate areas are identified by the novel structure based on deep belief network. Finally, the corresponding alarm is given by the identification results.
The algorithm schematic diagram is shown in Figure 2.

III. EXPERIMENTAL RESULTS

A. *Database*

The following simulations were done on a PC Processor Intel(R) Core (TM) i5-3337U CPU @ 1.80GHz, 1801 MHz, RAM 4GB. Some of the used frames extracted from videos in our database are presented in Figure 3.

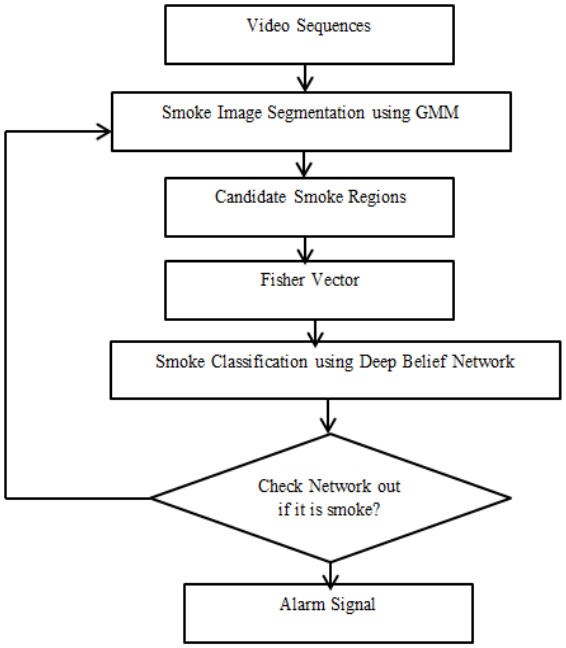

Figure 2. Algorithm Shematic Diagram

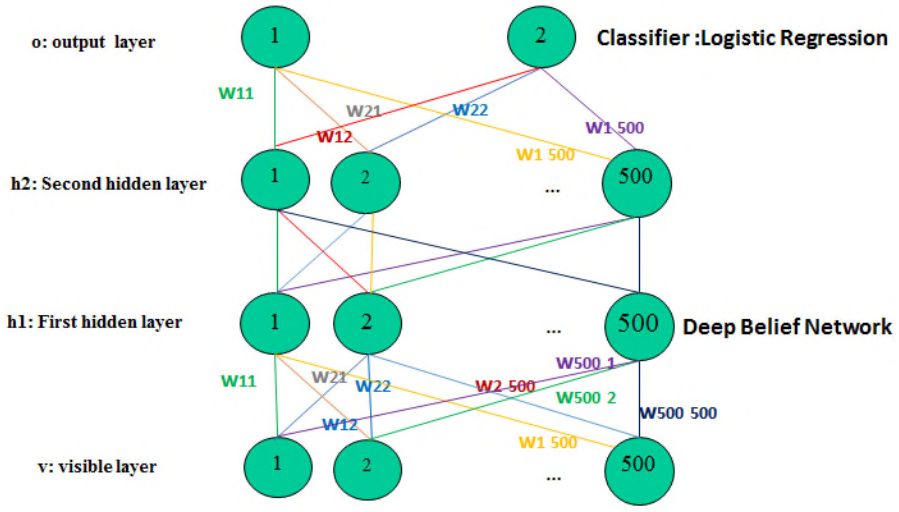

Figure 1. Deep Belief Network Architecture

B. Discussion

First of all, we have extracted frames from smoke-based videos. The size of each frame of the videos is set to 320 × 240 pixels. The main idea of our work is to extract the candidate regions related to the smoke movement which will inserted in a vector of characteristics.

The use of criterion of Fisher allowed us to keep the most important values in this feature vector to make the classification then by the deep belief network. As we mentioned before, we used the GMM [1] as a technique to detect the motion of smoke that is considered a chaotic movement. The results are shown in the Figure 4.

Figure 3. (a) Frames containing smoke extracted from our Database without noise, (b) Frames containing smoke with noise (fog, moving people, etc.), (c) Frames without smoke

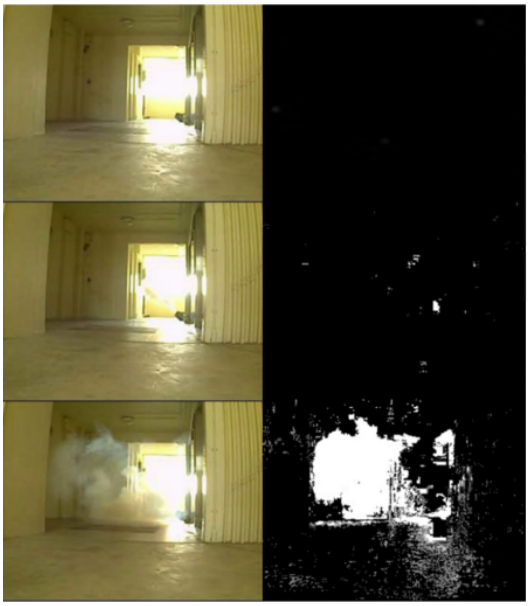

Figure 4. Background Substraction using GMM [1]

 ISBN: 978-1-61208-716-0

At this stage, we have extracted frames from smoke-based videos. The size of each frame of the videos is set to 320 × 240 pixels. The fixed parameters of the deep belief network are defined as:
Number of hidden layers=2
Number of epochs=400
Learning rate = 0.6

After fixing the features that we classified using deep belief network, we changed the classifier to SVM [7]. The comparison between the proposed method and the method using SVM is based on these criteria:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

$$Precision = \frac{TP}{TP + FP} \quad (5)$$

$$Recall = \frac{TP}{TP + FN} \quad (6)$$

$$F1score = 2.\frac{Precision.Recall}{Precision + Recall} \quad (7)$$

where TP: True Positive smoke Frames
TN: True Negative Frames
FP: False Positive Frames
FN: False Negative Frames

In this section, the frames containing noise such as moving people and fog give us an idea about the robustness of the proposed method. We notice a slight modification in the different criterion of comparison. These results are exhibited in the Table I.

This table shows that the developed method presents in the absence of the noise the best value of accuracy and Recall compared to the classification with SVM [7]. The presence of the noise affects slightly the accuracy and the precision. These values decrease slowly with noise. Moreover, the application of our proposed method helps us to find the smoke and the no smoke regions, as shown in Figure 5.

TABLE I. COMPARISON BETWWEN THE PROPOSED METHOD USING DBN AND THE METHOD USING SVM CLASSIFIER FOR SMOKE DETECTION

Condition	Classifier	Accuracy	F1 score	Precision	Recall
-Noise	**SVM**	93.43	0.952	1	0.912
	DBN	94.86	0.962	1	0.928
+Noise	**DBN**	92.57	0.936	0.88	1
	SVM	91.54	0.92	0.86	1

Figure 5. Smoke Detection Results

 ISBN: 978-1-61208-716-0

IV. CONCLUSION AND PERSPECTIVES

In this paper, we proposed a novel approach for smoke detection using Deep Belief Network. We extracted the candidate regions by using GMM [1].

For a better classification, we used the Fisher vector which gives us maximum separation between the means of different classes and keeps just the most important values. Then, the extracted feature vector was fed into deep belief network to calculate many criteria such as accuracy, F1 score, Precision and Recall. The robustness of this method is tested by adding noise. Finally, to evaluate the noise influence, we tested our method on noisy data. These promising results provide clear evidence about the capability of such a network in recognizing smokes in our recorded frames easily. Train the developed method with a significant amount of data will undoubtedly provide more promising results and would help in accelerating the next generation of surveillance and real-time monitoring systems. As perspectives to our work, the future step is to combine the deep belief network to FasterR-CNN to classify and localize simultaneously smoke and no smoke regions.

ACKNOWLEDGMENT

This work was supported by the Cooperation Project: PHC Utique 41755XB (CMCU 19G1126) and the CARTT-IUT: University of Toulon.

REFERENCES

[1] J. Yang, X. Yuan, X. Liao, P. Liul, G. Sapiro, and L. Cari, “Video Compressive Sensing Using Gaussian Mixture Models,” IEEE Transactions on Image Processing, 2014, pp. 4863-4878.

[2] Y. LeCun, Y. Bengio, and G .Hinton, “Deep Learning,” International Journal of Science Nature, 2015, pp. 436 - 444.

[3] G. Hinton, S. Osindero, and Y. Whye, “A fast learning algorithm for deep belief nets,” Neural Computation, Toronto, 2006, pp.1527-1554.

[4] Y. Hu and X. Lu, “Real-time video fire smoke detection by utilizing spatial-temporal ConvNet features,” Multimedia Tools and Applications, 2018, pp. 29283 – 29301.

[5] Z. Yin, B. Wan, F. Yuan, X. Xia, and J. Shi, “A Deep Normalization and Convolutional Neural Network for Image Smoke Detection,” IEEE Transactions on Digital Object Identifier, 2017, pp. 18429-18438.

[6] J. Chen, Y. He, and J. Wang, “Multi-feature fusion based fast video flame detection”, Build. Environ. 45 (May 2010) pp. 1113–1122.

[7] C. Hsu and C. Jen Lin, “A Comparison of Methods for Multiclass Support Vector Machines”, IEEE Transactions on neural networks, Taiwan 2002 , pp. 1–26.

[8] U. Toreyin, Y. Dedeoglu, and E. Cetin, “Contour Based Smoke Detection in Video Using Wavelets”, 14th European Signal Processing Conference (EUSIPCO 2006), Florence, Italy, 2006, pp. 1-5.

[9] S. Mikat, G. fitscht, J. Weston, B. Scholkopft, and K. Mullert, “Fisher discriminant analysis with kernels”, Neural Networks for Signal Processing IX, 1999, pp. 41 - 48.

[10] N. Zhangab, S. Dingab, J. Zhangab, and Y. Xuec, “An overview on Restricted Boltzmann Machines”, Neurocomputing, 2018, pp. 1186-1199.

[11] A. Khotanzad and C. Chung, “Application of Multi-Layer Perceptron Neural Networks to Vision Problems”, Neural Computing and Applications, London, 1998, 9, pp. 249 - 259.

[12] C. Chang and R. Chao, “Application of back-propagation networks in debris flow prediction”, Engineering Geology, 2006, pp. 270–280.

 ISBN: 978-1-61208-716-0

EFM-HOG: Improving Image Retrieval in the Wild

Sugata Banerji

Lake Forest College
555 North Sheridan Road
Lake Forest, IL 60045, USA
Email: banerji@lakeforest.edu

Atreyee Sinha

Edgewood College
1000 Edgewood College Drive
Madison, WI 53711, USA
Email: asinha@edgewood.edu

***Abstract*—The problem of retrieving images from a dataset, which are similar to a query image is an important high-level vision problem. Different tasks define similarity based on different low-level features like shape, color or texture. In the presented work, we focus on the problem of retrieval of images of similarly shaped objects, with the query being an object selected from a query image at runtime. Towards this end, we propose a novel shape representation and associated similarity measure, which exploits the dimensionality reduction and feature extraction methods of Principal Component Analysis (PCA) and Enhanced Fisher Model (EFM). The effectiveness of this representation is demonstrated on large-scale image datasets for the task of object retrieval and the performance is compared to Histograms of Oriented Gradients (HOG).**

Keywords–Computer Vision; Principal Component Analysis; Fisher Linear Discriminant; Enhanced Fisher Model; Histogram of Oriented Gradients; Image Search.

I. Introduction and Background

With the enormous popularity of digital devices equipped with cameras, along with the wide access to high speed Internet and cloud storage, several applications based on image search and retrieval have emerged. Such applications include augmented reality, geo-localization, security and defense, educational uses, to name a few. Billions of images are uploaded and shared over social media and web sharing platforms everyday, giving rise to a greater need for systems that can retrieve images similar to a query image from a dataset. Traditional approaches of content-based image retrieval are based upon low level cues such as shape, color and texture features. In this paper, we are trying to address the problem of retrieving images that have similarity in the shapes. Specifically, we select a window from a query image surrounding an object of interest and want to be able to retrieve similarly shaped objects from other images in the dataset, which are taken "in the wild", i.e., user generated content without any control. Towards that end, we investigate and propose a novel representation and retrieval technique that is based on shape features, dimensionality reduction and discriminant analysis and is robust to the slight changes in the window object selection.

The Histograms of Oriented Gradients (HOG) feature vector [1], originally proposed for pedestrian detection, is very popular among researchers for shape matching. It has successfully been combined with other techniques [2] and fused with other descriptors [3] for scene image classification. HOG has also given rise to other extremely successful object detection techniques, such as Deformable Part Models (DPM) [4]. More complicated descriptors [5] have been used for image retrieval with reasonable success. However, such methods are time consuming and more processor-intensive as compared to simple HOG matching. In recent years, handcrafted features have declined in popularity due to the success of deep neural networks in object recognition [6]–[8], but such methods are not without their drawbacks. Deep neural networks require a lot of processor time and run better on specialized hardware. They also require far greater number of training images than are available in a small or medium-sized dataset. For these reasons, enhancing simple handcrafted features like HOG can be effective for solving small-scale retrieval problems more effectively than more complex methods.

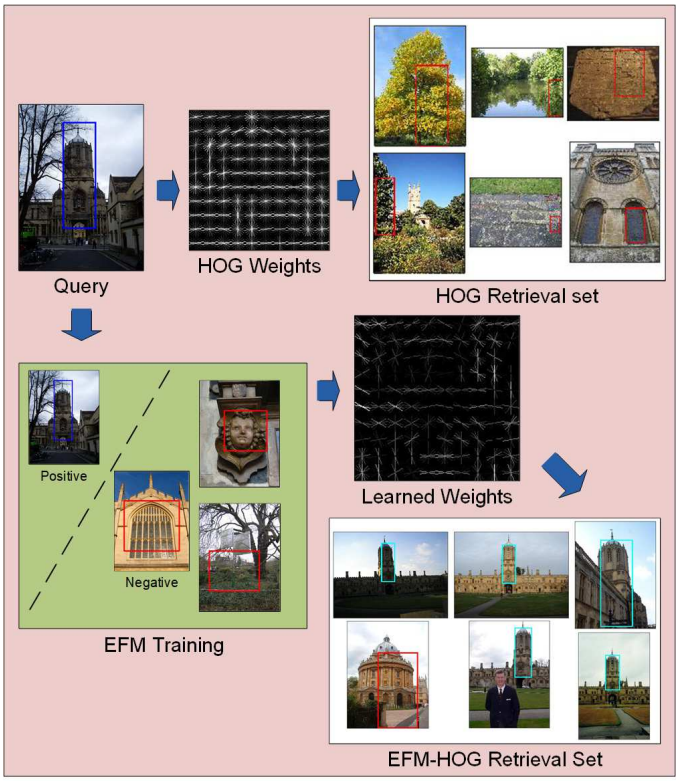

Figure 1. The proposed image representation aims at enhancing the HOG-based retrieval set by training an EFM-based classifier. The method is described in more detail in Section II

Simple HOG matching, however, poses significant challenges in effective image retrieval due to the fact that the apparent shape of the query object may change considerably

 ISBN: 978-1-61208-716-0

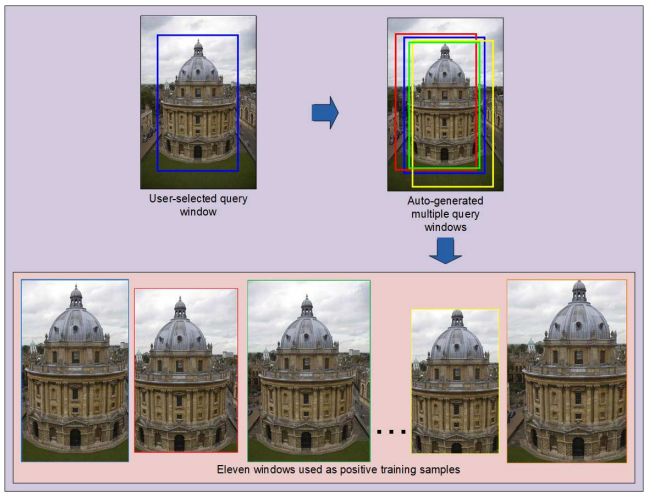

Figure 2. Auto-generation of offset windows to be used as positive training samples during querying. The window dimensions and offsets shown are only representative.

between images due to differences in lighting, viewing angle, scale and occlusion. This is particularly true for content generated by users in the wild. In effect, every query image is an exemplar of its own class and a retrieval system must be trained to treat it that way. In [9], this idea is handled using a Support Vector Machine (SVM) [10]. Instead of an SVM, here we introduce the novel idea of enhancing the HOG features by the EFM process [11] because it produces a low-dimensional representation, which is important from the computational aspect. Principal Component Analysis (PCA) has been widely used to perform dimensionality reduction for image indexing and retrieval [11]. The Enhanced Fisher Model (EFM) feature extraction method has achieved good success rates for the task of image classification and retrieval [3]. In the proposed method, which is represented schematically in Figure 1, we show this method to be effective in isolating the query object from the background.

The rest of this paper is organized as follows. Section II outlines in detail the method proposed in this paper. The datasets used and the experiments performed are detailed in Section III. Finally, we list our conclusions and directions for future research in Section IV.

II. Proposed Method

A. Window Generation

We start with generating objectness windows from each image. We use the method used by [12], which designs an objectness measure and explicitly trains it to distinguish windows containing an object from background windows. This method uses five objectness cues - namely, multi-scale saliency, color contrast, edge density, superpixels straddling, and location and size - and combines them in a Bayesian framework. We select the 25 highest-scoring windows from each image in our dataset and extract HOG features from these windows.

While testing our system, the user generates a window on the query image manually roughly enclosing the object of interest. Then, we automatically select 10 slightly offset versions of this window. Eight of these are generated by moving the user-selected window to the right, left, up, down, up-right, up-left, down-right and down-left by 5%, respectively. Two windows are generated by expanding and contracting the user's selection by 5%, respectively. Features are now extracted from these 10 as well as the original window for further processing. This process is represented in Figure 2.

B. HOG

The idea of HOG rests on the observation that local features such as object appearance and shape can often be characterized well by the distribution of local intensity gradients in the image [1]. HOG features are derived from an image based on a series of normalized local histograms of image gradient orientations in a dense grid [1]. The final HOG descriptors are formed by concatenating the normalized histograms from all the blocks into a single vector.

Figure 3 demonstrates the formation of the HOG vector for a window selected from an image. We use the HOG implementation in [13] for both generating the descriptors and rendering the visualizations used in this paper.

C. Dimensionality Reduction

PCA, which is the optimal feature extraction method in the sense of the mean-square-error, derives the most expressive features for signal and image representation. Specifically, let $\mathcal{X} \in \mathbb{R}^N$ be a random vector whose covariance matrix is defined as follows [14]:

$$S = \mathcal{E}\{[\mathcal{X} - \mathcal{E}(\mathcal{X})][\mathcal{X} - \mathcal{E}(\mathcal{X})]^t\} \quad (1)$$

where $\mathcal{E}(\cdot)$ represents expectation and t the transpose operation. The covariance matrix S is factorized as follows [14]:

$$S = \Phi \Lambda \Phi^t \quad (2)$$

where $\Phi = [\phi_1 \phi_2 \cdots \phi_N]$ is an orthogonal eigenvector matrix and

$$\Lambda = diag\{\lambda_1, \lambda_2, \ldots, \lambda_N\}$$

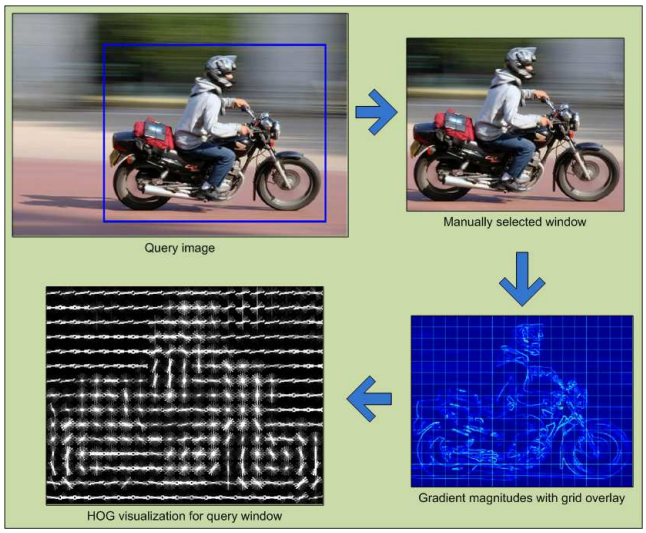

Figure 3. Formation of the HOG descriptor from a query image window.

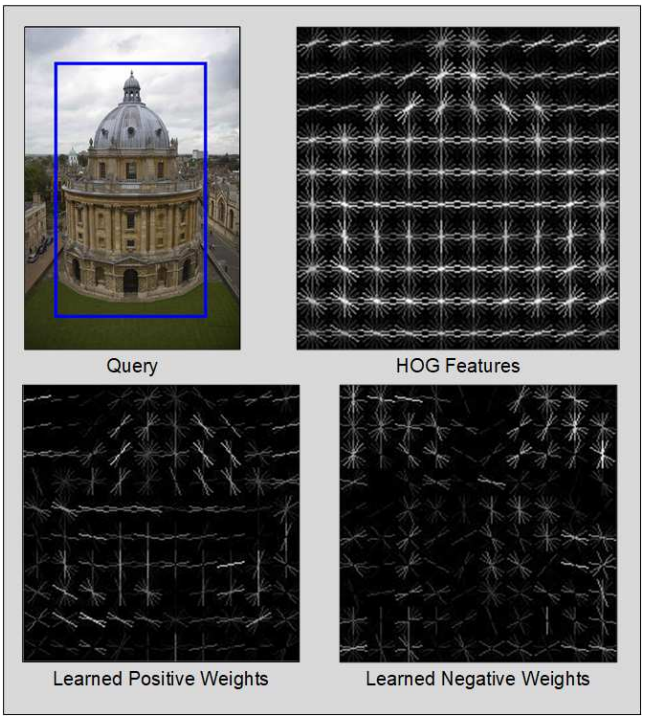

Figure 4. The positive and negative weights learned from the HOG features through the EFM discriminative feature extraction process.

a diagonal eigenvalue matrix with diagonal elements in decreasing order. An important application of PCA is the extraction of the most expressive features of $\mathcal{X}$. Towards that end, we define a new vector $\mathcal{Y}$: $\mathcal{Y} = P^t\mathcal{X}$, where $P = [\phi_1\phi_2 \dots \phi_K]$, and $K < N$. The most expressive features of $\mathcal{X}$ thus define the new vector $\mathcal{Y} \in \mathbb{R}^K$, which consists of the most significant principal components.

D. EFM

The features obtained after dimensionality reduction by PCA as discussed in Section II-C are the most expressive features for representation. However, they are not the optimum features for classification. Fisher's Linear Discriminant (FLD), a popular method in pattern recognition, first applies PCA for dimensionality reduction and then discriminant analysis for feature extraction. Discriminant analysis often optimizes a criterion based on the within-class and between-class scatter matrices S_w and S_b, which are defined as follows [14]:

$$S_w = \sum_{i=1}^{L} P(\omega_i)\mathcal{E}\{(\mathcal{Y} - M_i)(\mathcal{Y} - M_i)^t|\omega_i\} \quad (3)$$

$$S_b = \sum_{i=1}^{L} P(\omega_i)(M_i - M)(M_i - M)^t \quad (4)$$

where $P(\omega_i)$ is *a priori* probability, ω_i represent the classes, and M_i and M are the means of the classes and the grand mean, respectively. One discriminant analysis criterion is J_1: $J_1 = tr(S_w^{-1}S_b)$, and J_1 is maximized when Ψ contains the eigenvectors of the matrix $S_w^{-1}S_b$ [14]:

$$S_w^{-1}S_b\Psi = \Psi\Delta \quad (5)$$

where Ψ, Δ are the eigenvector and eigenvalue matrices of $S_w^{-1}S_b$, respectively. The discriminating features are defined by projecting the pattern vector $\mathcal{Y}$ onto the eigenvectors of Ψ:

$$\mathcal{Z} = \Psi^t\mathcal{Y} \quad (6)$$

$\mathcal{Z}$ thus contains the discriminating features for image classification.

The FLD method, however, often leads to overfitting when implemented in an inappropriate PCA space. To improve the generalization performance of the FLD method, a proper balance between two criteria should be maintained: the energy criterion for adequate image representation and the magnitude criterion for eliminating the small-valued trailing eigenvalues of the within-class scatter matrix. The EFM improves the generalization capability of the FLD method by decomposing the FLD procedure into a simultaneous diagonalization of the within-class and between-class scatter matrices [11]. The simultaneous diagonalization demonstrates that during whitening the eigenvalues of the within-class scatter matrix appear in the denominator. As shown by [11], the small eigenvalues tend to encode noise, and they cause the whitening step to fit for misleading variations, leading to poor generalization performance. To enhance performance, the EFM method preserves a proper balance between the need that the selected eigenvalues account for most of the spectral energy of the raw data (for representational adequacy), and the requirement that the eigenvalues of the within-class scatter matrix (in the reduced PCA space) are not too small (for better generalization performance). For this work the number of eigenvalues was empirically chosen.

E. Training

The EFM feature extraction method uses positive and negative training samples to find the most discriminative features.

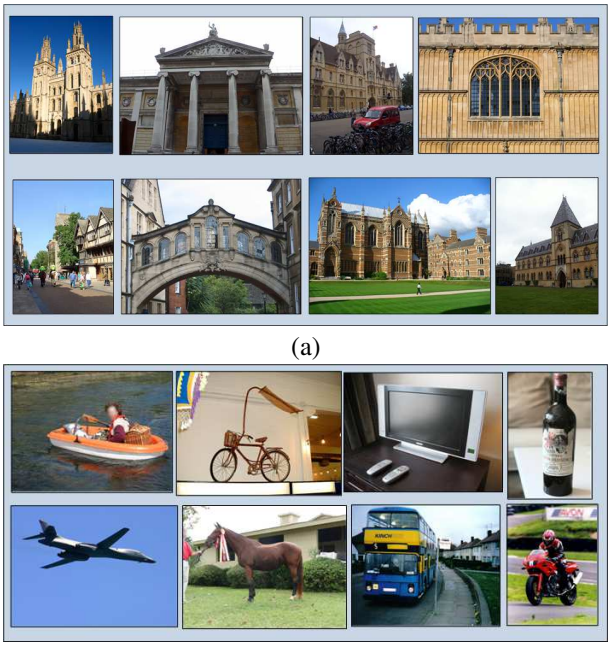

Figure 5. Some sample query images from (a) the Oxford Buildings dataset, and (b) the PASCAL VOC 2012 dataset.

In our setting, there is only one query image to be used as a positive sample. This is similar to the Exemplar-SVM training scenario used by [9], but to make the training more robust to selection error by the user and to prevent overfitting, we use 11 windows instead of just the one selected by the user as described in Section II-A.

We rank all objectness windows from all images in the dataset in terms of Euclidean distance in the HOG space from the original query window. For the negative training samples, we use 110 windows that are ranked low, i.e., are very distant in the HOG space. Experimentally, we found that the last ranked windows are not very good candidates for negative training samples, since they are often outlier windows that contain large blank areas like the sky. Instead, windows that have a rank 1000 to 5000 perform well. We also tried training the system with different numbers of negative samples and found a number close to 100 performs the best. These windows are mostly background regions like ground and vegetation. The positive and negative weights for the HOG features learned by this method can be seen in Figure 4.

For an n-class problem, the EFM process for discriminatory feature extraction reduces the dimensionality of any vector to $n-1$. Since our problem is a two-class problem, EFM produces one feature per window. We compute the score of each window by finding the absolute value of the difference between the window EFM feature and the average positive training set EFM feature. Ranking the images by their best-scoring windows gives us the retrieval set.

III. Experiments

A. Dataset

We have used the two datasets shown in Figure 5 for this work. First, we evaluate the retrieval performance of the proposed method on images gathered in the wild. For this, we use the Oxford Buildings dataset [15], which consists of 5062 images of 11 different Oxford landmarks and distractors collected from Flickr [16]. 55 images from this dataset were used as queries for testing our retrieval system. Flickr images are completely user-generated, which means there is a great variation in camera type, camera angle, scale and lighting conditions. This makes this dataset very difficult for image retrieval in general and landmark-identification in particular (the results of which are shown in Figure 6). Figure 5(a) shows some of our query images from this dataset. For each query, the images that contain the query landmark are further classified into *good*, *OK* and *junk* categories, with progressively poorer views of the query landmark. Table I shows the landmark-wise distribution of *good*, *OK* and *junk* images in this dataset.

TABLE I. The number of images containing each landmark in the Oxford Buildings dataset

Landmark	Good	OK	Junk
All Souls Oxford	24	54	33
Ashmolean Oxford	12	13	6
Balliol Oxford	5	7	6
Bodleian Oxford	13	11	6
Christ Church Oxford	51	27	55
Cornmarket Oxford	5	4	4
Hertford Oxford	35	19	7
Keble Oxford	6	1	4
Magdalen Oxford	13	41	49
Pitt Rivers Oxford	3	3	2
Radcliffe Camera Oxford	105	116	127

We also test retrieval performance on the PASCAL VOC 2012 dataset [17]. We only use the training/validation data from this dataset to test our retrieval algorithm. This data consists of 17,125 images from 20 classes. We create five random test sets of size 100 each from the original image set and perform a five-fold cross-validation on all our experiments. Figure 5(b) shows some images from this dataset.

B. The Retrieval Task

The proposed image representation is tested on two different tasks the first of which is retrieval. Here, an image is used as a query to retrieve similar scenes from the dataset. For this, the user selects a rectangular region of interest from the query image, and HOG features from this rectangular window is matched with the 25 highest scoring objectness windows from each image in the database, both in the raw HOG space and after the proposed training and feature extraction procedure. The closest matches based on Euclidean distance are retrieved

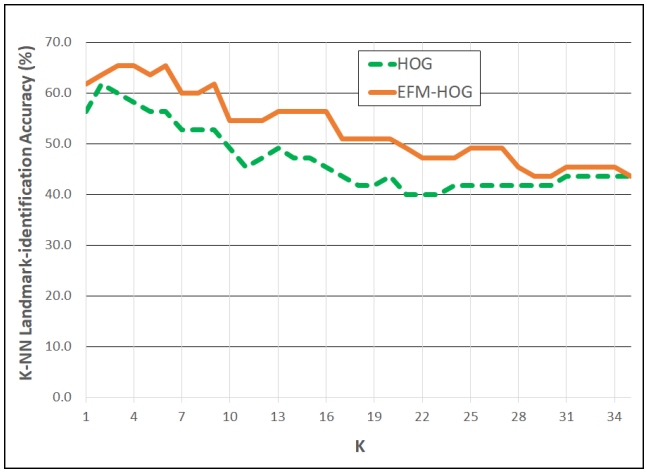

Figure 6. The mean landmark-identification performance by using the K-nearest neighbors method with varying K.

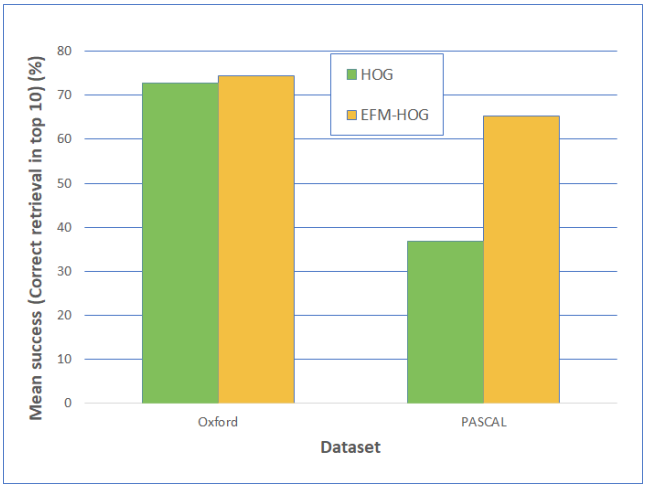

Figure 7. Mean retrieval accuracy (measured by the presence of a relevant image in the top 10 retrieved images).

 ISBN: 978-1-61208-716-0

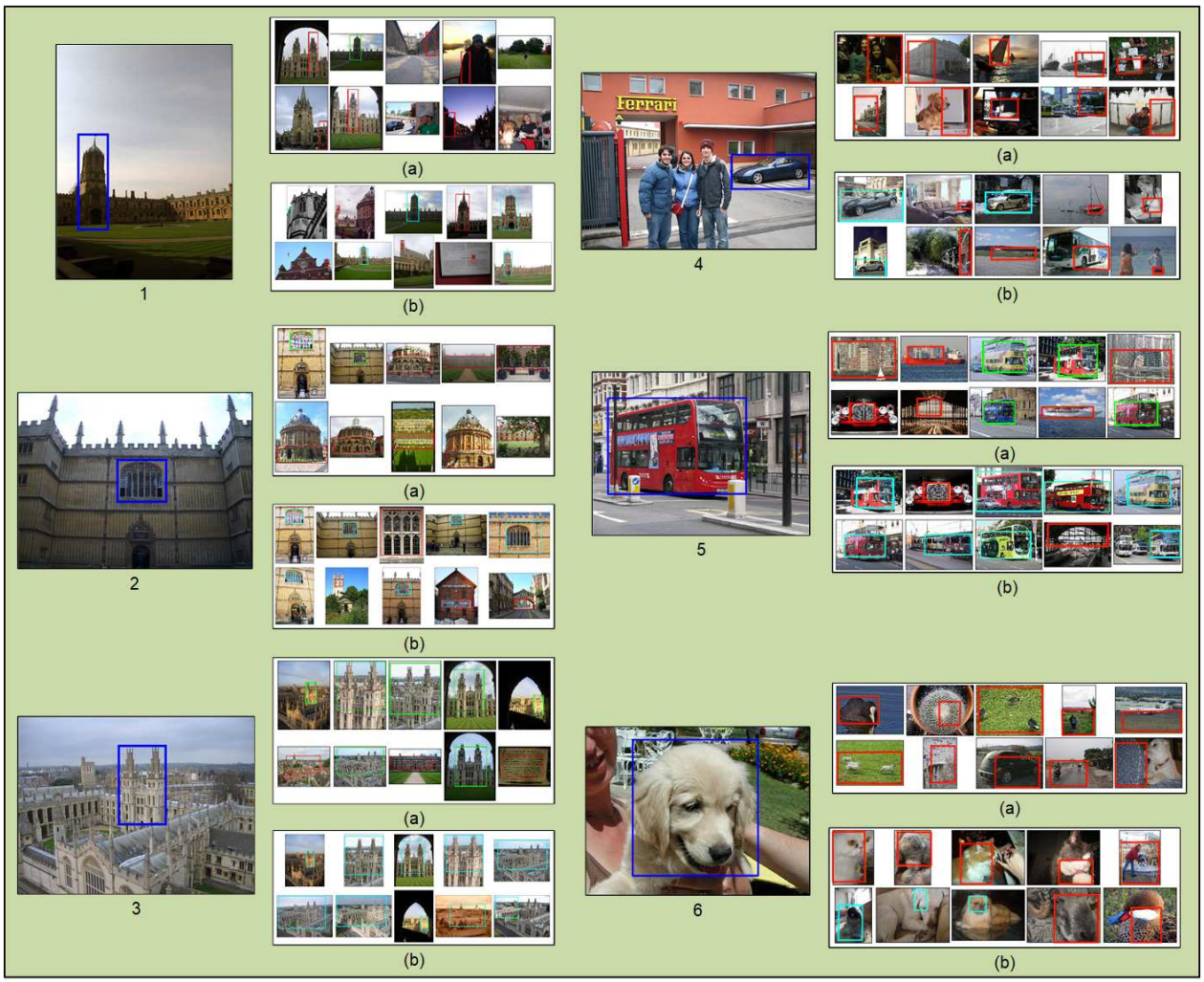

Figure 8. Comparison of image retrieval results for HOG and the proposed EFM-HOG. Images 1, 2 and 3 are from the Oxford Buildings dataset. Images 4, 5 and 6 are from the PASCAL VOC 2012 dataset. In each case, (a) shows top ten images retrieved by HOG, and (b) shows top ten images retrieved by EFM-HOG. Red rectangles indicate images that do not represent the same landmark or object class as the query.

in order of their distance from the query window. Finding an instance of the query in the top 10 retrieved images is considered a success. Figure 7 compares the retrieval success rates of the HOG descriptor and the proposed EFM-HOG representation. Specifically, in 41 cases out of 55 queries in the Oxford buildings dataset, the query landmark is retrieved within top 10 images by the proposed method, as opposed to 40 by HOG. This is actually a very small difference, but this can be explained by the nature of this dataset. For all landmark query images in this dataset, there are at least some images in the dataset that show clear views of the landmarks with no occlusions. HOG is actually pretty effective at retrieving these images. To actually understand the effectiveness of the proposed method, we repeat this experiment with just the *junk* files for each query. In this experiment, we find that the HOG method retrieves a relevant image in the top 10 only once out of all 55 queries, while the proposed EFM-HOG method achieves this 5 times out of the 55.

For PASCAL VOC, the experiment is performed on all five random splits and the average success rate is found to be 65.2% for EFM-HOG as compared to 36.8% for HOG. We also find that the conventional HOG performs quite well for clearly segmented objects, such as airplanes in the sky, but the EFM-HOG performs much better for images of objects with a cluttered background. Some HOG and EFM-HOG retrieval results are shown in Figure 8. Figure 9 shows another interesting aspect of our retrieval technique. Here, we show the image means of the first 100 windows retrieved by both HOG and EFM-HOG on PASCAL VOC. The figure shows that the EFM-HOG means contain clearer shapes, which indicates that the EFM-HOG retrieves more similar shapes than HOG, even when the results are irrelevant to the query.

C. The Landmark-identification Task

Some images in our Oxford Buildings dataset belong to one of the eleven landmarks listed in Table I, the others belong to none of the classes and are used as distractors. The second experiment that we performed with the new EFM-HOG descriptor was a landmark-identification task where the system tries to label each query image with its correct landmark label. This is done by retrieving relevant images in a manner similar to the retrieval task, and then performing the K-

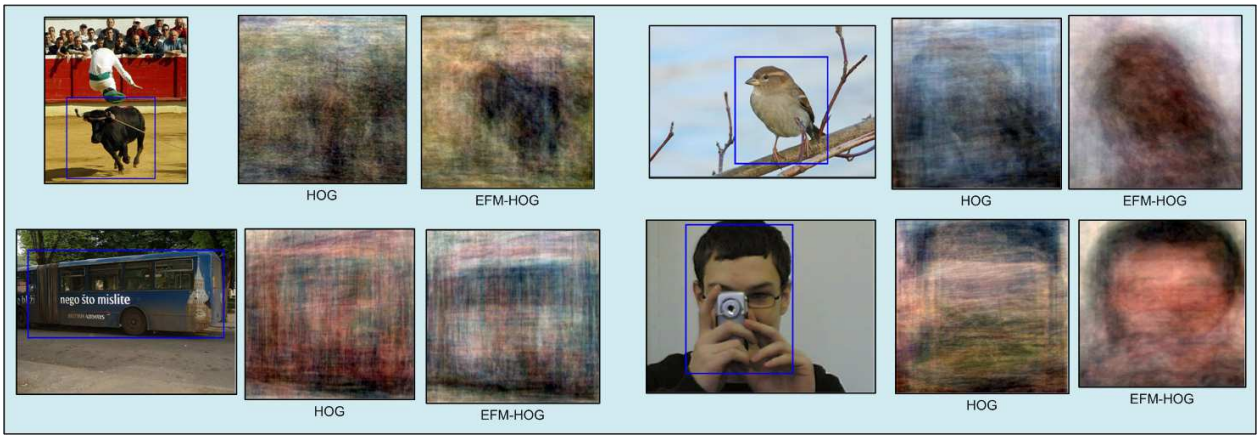

Figure 9. The means of the top 100 retrieved windows for HOG and EFM-HOG for 4 query images from the PASCAL VOC 2012 dataset.

nearest neighbors classification on the top K results. The same experiments are repeated for the conventional HOG descriptor as well. As can be seen from Figure 6, the proposed EFM-HOG outperforms HOG by a significant margin for nearly all values of K between 1 and 35. The highest EFM-HOG landmark-recognition performance of 65.5% is achieved at $K = 3$.

IV. Conclusion

We have presented in this paper a new image descriptor based on HOG and discriminant analysis that uses a novel approach to fetch scenes with similar shaped objects. We have conducted experiments using over $5,000$ images from the Oxford Buildings dataset and over $17,000$ images from the PASCAL VOC 2012 dataset and concluded the following: (i) HOG features are not always sufficiently discriminative to perform meaningful retrieval, (ii) the discriminative nature of HOG features can be improved with the EFM for feature extraction and dimensionality reduction, and (iii) HOG features perform well for clearly isolated objects with little background clutter, but the EFM-HOG performs better for real-world images with cluttered backgrounds.

We intend to use this method with more datasets in the future, so that a more thorough understanding of its strengths and weaknesses can be achieved.

Acknowledgment

The authors would like to thank Professor Jana Košecká at the Department of Computer Science, George Mason University, Fairfax, Virginia for some valuable input on the EFM-HOG method and the experiments conducted.

References

[1] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2005, pp. 886–893.

[2] S. Banerji, A. Sinha, and C. Liu, "Scene Image Classification: Some Novel Descriptors," in Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, 2012, pp. 2294–2299.

[3] A. Sinha, S. Banerji, and C. Liu, "Novel Color Gabor-LBP-PHOG (GLP) Descriptors for Object and Scene Image Classification," in Proceedings of the Eighth Indian Conference on Computer Vision, Graphics and Image Processing, ser. ICVGIP '12. ACM, 2012, pp. 58:1–58:8.

[4] P. F. Felzenszwalb, R. B. Girshick, D. A. McAllester, and D. Ramanan, "Object Detection with Discriminatively Trained Part-Based Models," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, no. 9, 2010, pp. 1627–1645.

[5] K. E. A. v. d. Sande, C. G. M. Snoek, and A. W. M. Smeulders, "Fisher and VLAD with FLAIR," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, June 2014, pp. 2377–2384.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in Proceedings of the Twenty-sixth Conference on Neural Information Processing Systems, 2012, pp. 1106–1114.

[7] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in Proceedings of the 3rd International Conference on Learning Representations, 2015.

[8] C. Szegedy, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going Deeper with Convolutions," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, 2015, pp. 1–9.

[9] T. Malisiewicz, A. Gupta, and A. A. Efros, "Ensemble of Exemplar-SVMs for Object Detection and Beyond," in Proceedings of the International Conference on Computer Vision, 2011, pp. 89–96.

[10] V. Vapnik, The Nature of Statistical Learning Theory. Springer-Verlag, 1995.

[11] C. Liu and H. Wechsler, "Robust Coding Schemes for Indexing and Retrieval from Large Face Databases," IEEE Transactions on Image Processing, vol. 9, no. 1, 2000, pp. 132–137.

[12] B. Alexe, T. Deselaers, and V. Ferrari, "Measuring the Objectness of Image Windows," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, no. 11, Nov 2012, pp. 2189–2202.

[13] A. Vedaldi and B. Fulkerson, "VLFeat: An open and portable library of computer vision algorithms," 2008 [accessed 2019-04-18].

[14] K. Fukunaga, Introduction to Statistical Pattern Recognition, 2nd ed. Academic Press, 1990.

[15] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object Retrieval with Large Vocabularies and Fast Spatial Matching," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2007, pp. 1–8.

[16] "Flickr," http://www.flickr.com, 2004, [accessed 2019-04-18].

[17] M. Everingham, L. J. V. Gool, C. K. I. Williams, J. M. Winn, and A. Zisserman, "The Pascal Visual Object Classes (VOC) Challenge," International Journal of Computer Vision, vol. 88, no. 2, 2010, pp. 303–338.

 ISBN: 978-1-61208-716-0

Study of Spectra with Low-Quality Resonance Peaks

Viktor A. Sydoruk
Institute of Bio- and Geosciences, IBG-2: Plant Sciences
Forschungszentrum Jülich GmbH
52428 Jülich, Germany
e-mail: v.sydoruk@fz-juelich.de

***Abstract*—A Python-based software for the phenoCAVe family of resonators for plants investigations is developed to automatically extract both the center frequency and the quality factor of the main resonance peak, i.e., of the lowest transversal-magnetic mode TM_{010}, at different scan positions. Due to the specific design of the cavities, which includes large openings on the top and on the bottom of the resonators, the main peak even in the unloaded case (when no object is measured) has a sufficiently small quality factor (<350), which leads to the large influence of the higher modes on the reliability of the extracted data. Additionally, the repeated movements of the resonators and long cables usage may alter the cable influence. Moreover, continuous movements during the scans as well as a finite time of spectra sweep give a distorted peak, especially at the borders of large objects, such as plant pots, i.e., when the objects are starting to pass through a resonator. All these problems and more are taken into account in the automatic data analysis software, which allows to obtain reliable responses from noninvasive scans of the investigated plants over the whole period of their growth and further usage of these responses to calculate the important parameters for the plant growth, such as water content, dry weight, biomass, pot water content etc. Here, the whole process from the analysis of distorted spectra to the evaluation of the suitable parameters for plant growth is demonstrated. Such an analysis can be used not only for plant phenotyping platforms but also in various physical platforms supposing low-quality spectra analysis and dielectric studies of materials.**

Keywords-low-quality resonance peaks; spectra fitting.

I. INTRODUCTION

Fitting of experimental data using theoretical models is an important and well-known problem in various fields of study allowing to fulfill the understanding of investigated processes or events. There are plenty of developed tools to do so, especially when theoretical models are quite simple, such as polynomial, *Gaussian*, and *Lorentzian*, or even a convolution of the last two, often called a *Voigtian*. Such tools, for instance Origin (OriginLab Corporation, Northampton, MA, USA), Matlab-based (MathWorks, Natick, MA, USA) scripts, Python-based scripts, such as the lmfit package [1], etc., are usually used in peak fitting tasks. Nevertheless, task-specified algorithms were developed as well for the peak fitting procedure, which mainly deal with either specific data or particular cases of some processes [3]–[6].

To analyze spectra where low-quality resonance peaks are presented, a simple model of fitting these peaks cannot be used due to several reasons. First, due to the coupling of modes, they are influencing each other by changing their visible parameters on the spectrum, such as peak frequency, peak amplitude/attenuation, as well as their quality factor. Secondly, the amplitude/attenuation at each frequency can be affected by uncertainties caused by reflections in cables, noises, spurious coupling or, in case of a phenoCAVe family of resonators [2], by continuous movements of the cavities. Some of these uncertainties can be represented as baselines which are fitted together with spectra to obtain proper spectrum parameters [7][8].

Skresanov et al. [9] described a novel approach to recover coupled mode parameters from the microwave resonator amplitude-frequency response to deal with the first reason mentioned above. They used an approach from the theory of oscillation, meaning that it is always possible to select such a coordinate system, in which oscillations are independent. In this case, the total amplitude $\Gamma(f)$ of the reflected/transmitted signal can be presented as a sum of complex amplitudes of these oscillations (modes) as follows:

$$\Gamma(f) = \Gamma_S + \sum_i^n \frac{A_i \exp(j\varphi_i)}{1 + 2jQ_i\tau_i(f)}, \quad (1)$$

where Γ_S is the coefficient equal to $\Gamma(f)$ at $f \to \infty$ and is considered to be a real number, i is the ordinal number of an oscillation, n is the total number of oscillations (modes), $\tau_i(f) = (f - f_{0i}) / f_{0i}$ is the frequency tuning parameter, A_i, Q_i, f_{0i}, and φ_i are the amplitude, quality factor, resonant frequency (Center Frequency, CF [2]), and phase shift of the i-th oscillation, respectively. In the case of scattering parameters, the amplitude $\Gamma(f)$ should be considered in the logarithmic form

$$S(f) = 20\log_{10}(|\Gamma(f)| / \Gamma_0), \quad (2)$$

where $S(f)$ is the scattering parameter, e.g., S_{11}, S_{21}, etc., and Γ_0 is the amplitude of the excited signal by a signal generator.

This work presents an advanced fitting algorithm based on the approach mentioned above [9] to study low-quality resonance peaks, i.e., the peaks that have low-quality factor, disturbances, influences of other modes, etc. The developed software allows to fully-automatically analyze all the spectra

 ISBN: 978-1-61208-716-0

obtained during the scans of a set of plants, which can be grown in different environmental conditions, may have different genotypes, pot sizes, soil types, etc.

In Section II, the selection of a proper approach to fit spectra obtained during the measurements by the phenoCAVe family of resonators [2] is presented, where the advantage of using (1) is clearly demonstrated (Section II.A) and compared with the simple *Lorentzian* fitting approach for the case of loaded resonators (Sections II.B and II.C). Additional problems that may appear during the fitting process caused by disturbed spectra are discussed in Section III. The spectra fitting routine and the graphical user interface of the fully-automated data analysis software based on it is shown in Section IV. The conclusions are conducted in Section V.

II. FITTING OF SPECTRA

A. Unloaded resonators

The fitting algorithm (further referred to as *complex* approach) that uses (1) to calculate parameters of up to 7 resonance modes for measured spectra is realized in so-called "Shaman" software [9]. As an example, Figure 1 shows the spectrum (black dots) which includes the resonance peaks of 5 different Transversal-Magnetic modes (TM_{010}, TM_{110}, TM_{210}, TM_{020}, and TM_{310}) as well as the fitted curve given by the software (red solid line). The obtained parameters for each mode are mentioned in TABLE 1. The matching is very close to ideal, which proves the right approach of the fitting. In many cases, especially when resonance peaks have large quality factors, the phase is not taken into account. The green dashed lines (Figure 1) are the fitted *Lorentzian* curves when each φ_i equal to zero. The difference is clearly visible, although the peaks have fairly correct positions.

B. Loaded resonators; comparison of fitting approaches

The approach when the phase is not taken into account can be realized in such a way that the data points are preselected to be as close as possible to the peak, i.e., in terms of the scattering parameter S_{21} up to 3-5 dB far from the highest point. Then, the fitting by *Lorentzian* curve is done in the following form (further referred to as *Lorentzian* approach)

$$\Gamma(f) = \frac{A}{\sqrt{1+4Q^2\tau^2(f)}}. \quad (3)$$

The last works well for the resonant oscillations, which quality factors are large enough, usually more than 10^3, to make the influence of other resonance modes negligible. Due to both the design and the large openings in the resonators used for plant investigations [2], the quality factor of the first (TM_{010}) mode even when the cavities are empty (unloaded) has far smaller values (< 350).

To demonstrate how a large error can be caused by the *Lorentzian* approach in comparison with the *complex* one, a study using both approaches was done. The analyzed spectra were obtained during the scan of a young maize plant (of about 2 g fresh weight) with its pot using resonator 1 [2]. The relative errors made by the *Lorentzian* approach are shown in Figure 2. Here, it was supposed that the *complex* approach gives the true values for the peak parameters, and the relative error was calculated using following equation

$$\text{Relative error} = \left| p_{complex} - p_{Lorentzian} \right| / \, p_{complex} \times 100\% \,, \quad (4)$$

where p denotes either peak frequency f_0 or quality factor Q. The result consists of the analysis of the 1st resonance peak only, although the fittings done by the *complex* approach included the 2nd mode too. It should be noted that the quality factor Q is the most suitable parameter in this study to demonstrate the influence of modes on each other, due to its strong dependence on the form of a resonance peak. Therefore, the relative errors were built versus Q on the plot. Decreasing of Q was stimulated by placing the resonator at different vertical positions during the scan of a maize plant with its pot [2].

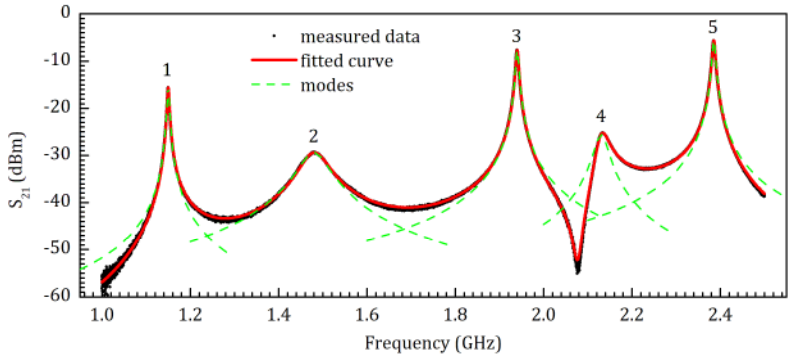

Figure 1. Transmittance spectrum with 5 resonance modes measured for unloaded resonator 1 [2] (black dots), fitting of it (red solid line), and individual resonance modes without phase shift φ_i information (green dashed line).

TABLE I. FITTED RESONANCE MODES' PARAMETERS

Modes, *i*	Parameters			
	f_{0i} (GHz)	Q_i	A_i (dBm)	φ_i
1: TM_{010}	1.14989	246.3	-15.55	0
2: TM_{110}	1.47970	23.2	-29.37	161°
3: TM_{210}	1.94015	297.4	-7.72	-5°
4: TM_{020}	2.13039	72.9	-25.60	8°
5: TM_{310}	2.38543	335.7	-5.68	-172°

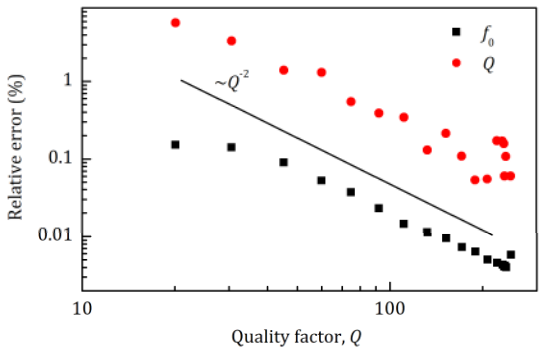

Figure 2. Relative errors of the estimated parameters, peak frequency f_0 (black squares) and quality factor Q (red circles), versus Q for the 1st resonance mode TM_{010}. The black solid line depicts the inverse square dependence on Q.

C. *The false approach leads to errors in plant water amount estimation*

It is visible that the difference between approaches may cause errors in the estimation of both f_0 and Q of up to 0.2% and 6%, respectively (see Figure 2). These errors may increase or decrease for larger or smaller measured plants, respectively. Moreover, by further estimation of the plant Water Amount (WA) [2] 0.2% by f_0 means about 2.3 MHz for the resonator 1, which in terms of the water distribution over the height of a plant gives ~1.5 µL/mm (data not published yet). This in its turn, for a young maize plant of 2 g fresh weight and about 200 mm height gives 0.3 mL of WA, or about 16% of error for the plant WA estimation (0.3[mL] / (2[g]×0.95[mL/g]) × 100% ≈ 16%)

This example demonstrates how the wrong fitting approach can affect the final measurement results. Therefore, in the phenoCAVe data analysis software, the *complex* approach instead of the simpler *Lorentzian* one was selected.

III. DISTURBED SPECTRA

A. *When a resonator is continuously moving while receiving the spectra*

Scanning of a plant by a resonator [2] involves the obtaining of a set of spectra at different positions along the height of the plant. Either a plant or a resonator can be moved against each other. The measurement setup based on the resonator 1 is supposed to shift a plant through the cavity, when the setup based on the resonator 2 displaces the cavity itself. To decrease the scanning time, these movements can be continuous at the intermediate points between the highest and the lowest positions. These lead to additional uncertainties caused by the finite sweep time of the Vector Network Analyzers (VNAs), which are used to obtain spectra, i.e., each frequency on the spectrum has its own position.

For the Screen-House setup [2], every position of the resonator is read out from the MAXON motor drive unit each 50 ms. The VNA (ZNC 3, Rohde & Schwarz GmbH, Cologne, Germany), in its case, has 87.5 µs sweep time per frequency point, i.e., one whole spectrum that consists of 801 points is measured during 70 ms. The highest speed of the resonator varies from 20 to 70 mm/s, meaning that during 70 ms the resonator can change its position by about 5 mm, which may lead to the crucial modification of the spectra, especially when the resonator is close to the plant pot. By knowing the vertical position of each point on the spectra and taking a set of spectra measured at different positions, spectra for each selected position can be recalculated using a polynomial fit of order 2. Such a polynomial fit gives a few positive effects. The first one is the automatically smoothed spectrum, and the second one is the possibility to detect outliers in combination with the Grubbs' test [10].

B. *Influence of cables and surroundings*

Both setups presented previously [2] have cables to deliver and acquire signals. The cables can be calibrated using a built-in utility of VNA and commercially available calibration kits (in our case ZV-Z132, Rohde & Schwarz GmbH, Cologne, Germany). The calibration may not properly work in some cases. For the setup based on resonator 2, even calibrated VNA showed a sinusoidal behavior on the spectrum (see Figure 3). In this case, additionally to the spectrum fitting, residuals $S_{21}^{res}(f)$ should be fitted by using a sum of sines in the following form

$$S_{21}^{res}(f) = \sum_{i}^{k} a_{i\sin} \sin(2\pi f / f_{i\sin} + \varphi_{i\sin}), \qquad (5)$$

where $a_{i\sin}$, $f_{i\sin}$, $\varphi_{i\sin}$ are the sine parameters, and k is the number of sines. The bottom plot in Figure 3 reflects this situation. Residuals may depend on the position of the resonator due to the influence of surroundings, causing the residuals fitting to be done for each spectrum. Moreover, this procedure should be circled to obtain proper parameters for the resonance peak, i.e., firstly, the spectrum (measured data)

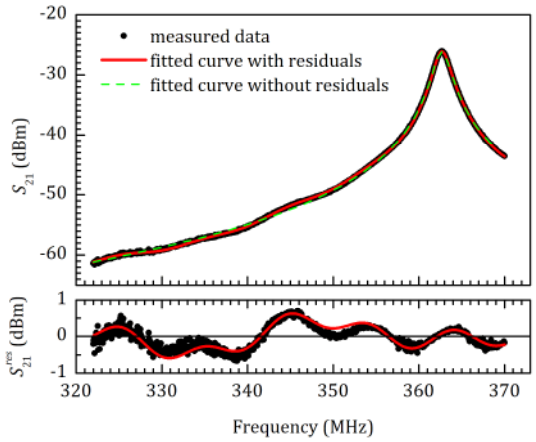

Figure 3. *Complex* fitting approach for the spectrum obtained using unloaded resonator 2 [2] (top plot) and fitting of residuals S_{21}^{res} using (4) with $k = 2$.

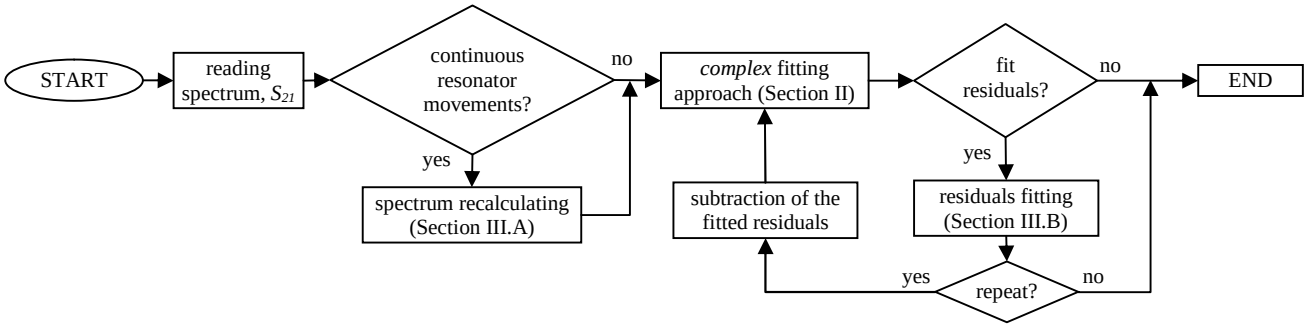

Figure 4. Suggested spectrum fitting routine.

is fitted by the *complex* approach, secondly, residuals are calculated and fitted, and finally, the obtained curve in the form (5) is subtracted from the measured data. These three steps are repeated several times to obtain a stable solution (see Figure 3).

IV. SPECTRUM FITTING ROUTINE

The suggested spectrum fitting routine is shown in Figure 4. The whole procedure starts from the reading of the spectral information (measured data) for the resonance peak at the selected position of a resonator. Then, if the resonator was continuously moving during the scan, the recalculating spectrum function is called (Section III.A). After that, the *complex* fitting approach starts (Section II), which is followed by the residuals fitting if needed (Section III.B). After subtraction of the fitted residuals, the *complex* approach can be called again, i.e., circled until obtaining a stable solution. At the end, the evaluated parameters of the resonance peak are stored for further analysis (not a part of this work). This routine is repeated for each measured spectrum at different positions of the resonator.

It should be noted that the residuals obtained at different positions are stored too for the unloaded runs of the resonator. Later, these residuals are used to simplify the fittings for the resonator runs with a measured object, e.g., a plant, a pot with soil, etc.

The fitting routine presented in Figure 4 was programmed in the fully-automatic phenoCAVe data analysis software using Python programming language. To make an interface and to deal with the mathematics behind it, a list of packages was used, such as pyqtgraph, PyQt5, scipy, numpy, lmfit, csv, itertools, fnmatch, inspect, os, bayeos, multiprocessing, sys, warnings, time, traceback, typing, copy, etc. The software tab, where spectra fitting is visible, is presented in Figure 5 with the data shown in Figure 3. The time spent to fit one spectrum from "Start" to "End" (see Figure 4) using "spectrum recalculating" and 20 iterations of "residuals fitting" was less than 5 seconds on the Intel Xeon E5-2630 v3 based computer.

Figure 5. phenoCAVe data analysis software developed using Python programming language.

V. CONCLUSIONS

The fitting approach presented in this work was mainly developed to analyze low-quality resonance peaks on measured spectra. The simple *Lorentzian* fitting approach was compared with the *complex* one, to which a preference was given. Additionally to that, issues that may arise during the measurements using partially opened resonators were shown and discussed with their possible resolution. Among them are the influences of other modes, surroundings, cables, and continuous movements of either a resonator or an investigated object. In the end, the spectrum fitting routine was suggested and the software based on it was demonstrated. The suggested *complex* fitting approach is not newly developed but the proposed fully automated fitting routine has novel ideas which can be useful for the precise analysis of spectra with low-quality resonance peaks.

ACKNOWLEDGMENT

The author would like to thank A. Gubin with IRE, Kharkiv, Ukraine and I. Zadorozhnyi with ICS-8, Forschungszentrum Jülich GmbH, Germany for valuable discussions related to the fitting approach and Python-programming issues.

REFERENCES

[1] M. Newville et al., "LMFIT: Non-Linear Least-Squares Minimization and Curve-Fitting for Python," https://lmfit.github.io/lmfit-py/, retrieved: Apr. 2019.

[2] V. A. Sydoruk et al., "Design and Characterization of Microwave Cavity Resonators for Noninvasive Monitoring of

Plant Water Distribution," IEEE Trans. Microw. Theory Tech., vol. 64, no. 9, Sep. 2016, pp. 2894–2904, doi: 10.1109/TMTT.2016.2594218.

[3] T. Žák and Y. Jirásková, "CONFIT: Mössbauer spectra fitting program," Surf. Interface Anal., vol. 38, no. 4, Mar. 2006, pp. 710–714, doi: 10.1002/sia.2285.

[4] A. E. Hughes and B. A. Sexton, "Curve fitting XPS spectra," J. Electron Spectrosc. Relat. Phenom., vol. 46, no. 1, 1988, pp. 31–42, doi: 10.1016/0368-2048(88)80003-3.

[5] A. P. Hammersley and C. Riekel, "MFIT: Multiple spectra fitting program," Synchrotron Radiat. News, vol. 2, no. 1, Jan. 1989, pp. 24–26, doi: 10.1080/08940888908261193.

[6] A. J. Brown, "Spectral curve fitting for automatic hyperspectral data analysis," IEEE Transactions on Geoscience and Remote Sensing, vol. 44, no. 6, Jun. 2006, pp. 1601–1608, doi: 10.1109/TGRS.2006.870435.

[7] Q. Han et al., "Simultaneous spectrum fitting and baseline correction using sparse representation," The Analyst, vol. 142, no. 13, May 2017, pp. 2460–2468, doi: 10.1039/C6AN02341J.

[8] H. Hu et al., "Improved Baseline Correction Method Based on Polynomial Fitting for Raman Spectroscopy," Photonic Sensors, vol. 8, no. 4, Dec. 2018, pp. 332–340, doi: 10.1007/s13320-018-0512-y.

[9] V. N. Skresanov et al., "The novel approach to coupled mode parameters recovery from microwave resonator amplitude-frequency response," 2011 41st European Microwave Conference (EuMC 2011), IEEE Press, Oct. 2011, pp. 826–829, doi: 10.23919/EuMC.2011.6101922.

[10] "NIST/SEMATECH e-Handbook of Statistical Methods," http://www.itl.nist.gov/div898/handbook/, retrieved: Apr. 2019.

 ISBN: 978-1-61208-716-0

Residual Dense Generative Adversarial Network for Single Image Super-Resolution

Jiahao Meng
School of Science
Beijing University of Posts and Telecommunications
Beijing, P.R. China
Email: jiahaomeng@bupt.edu.cn

Zekuan Yu
Department of Biomedical Engineering
College of Engineering
Peking University
Beijing, P.R. China
Email: yuzekuan518@163.com

Tianping Shuai
School of Science
Beijing University of Posts and Telecommunications
Beijing, P.R. China
Email: tpshuai@bupt.edu.cn

***Abstract*—Model-based very deep Convolutional Neural Networks (CNN) have achieved great success in Single Image Super-Resolution (SISR) work. However, most of the super-resolution models based on deep convolution networks can not fully utilize the hierarchical features of the original low-resolution images. In order to improve the quality of the high-frequency details of the reconstructed super-resolution image, we proposes a super-resolution method for Residual Dense Generative Adversarial Networks (RDGAN). We use the Generative Adversarial Networks (GAN) as our main model structure and the residual-dense block as the basic building blocks of the generator, which makes the network pay more attention to the extraction of low-resolution image hierarchical features. Then, we fully exploit the hierarchical features from all the convolutional layers. Finally, we use perceptual loss as our loss function to get finer texture details and more realistic photo effects. Experiments show that our method can achieve significant improvement in the quality of high-frequency detail reconstruction at high magnification.**

Keywords–CNN; Single Image Super-Resolution; Generative Adversarial Networks.

I. INTRODUCTION

The task of estimating a High-Resolution (HR) image from its Low-Resolution (LR) counterpart is called Single Image Super-Resolution (SISR) [1], which has received significant attention and progress in recent years. Super-Resolution (SR) has direct applications in computer vision, such as image/video enhancement, medical image processing [2][3], face recognition [4] and image generation [5].

Image SR is an ill-posed problem. The ill-posed character of the under-constrained SR problem is especially pronounced for high upscaling factors. Recently, a large number of SISR methods have been proposed to solve this underdetermined problem, including interpolation-based [3][6], reconstruction methods [7], and learning-based methods [8][9]. Most CNN-based methods [10][11] attempt to minimize pixel-wise the Mean Square Error (MSE) between the ground truth image and the reconstructed HR image. This strategy calculates the pixel-wise image difference and maximizes the Peak Signal-to-Noise Ratio (PSNR), which is a common measurement for evaluating the SR algorithm. In these cases, the high-frequency details of some sharp edges and textures in the SR image are still blurred and smooth in appearance, which is significantly different from the ground truth image.

In order to solve these drawbacks, Ledig *et al.* [12] proposed a GAN-based network. This enhances the invariance of the pixel field change. However, for a very deep network, only using Residual Networks (ResNets) and jump connections can not fully utilize the LR image information. Inspired by Zhang *et al.* [13], we use Residual Dense Block (RDB) as the basic component of our generator and we use Local Residual Learning (LRL) in order to make full use of the hierarchical features of LR. In this paper, we proposes a deep learning SISR method, which uses enhance Residual Dense Generative Adversarial Network (RDGAN) to improve the reconstruction quality of high-frequency edges and textures in the SR images. At the same time, we minimize the perceptual loss so that the generated images have photo realistic textures.

The rest of this paper is organized as follows. Section II addresses the related works in the literature. Section III describes the method. Section IV describes the experiments. We conclude the paper in Section V.

II. RELATED WORKS

In order to solve the Single Image Super-Resolution problem, early algorithms [14][15] have been mainly based on sampling interpolation techniques, but these methods show considerable limitations in predicting the texture details of the image.

Recently, the CNN-based [16] approaches have shown excellent performance. Dong *et al.* proposed Super-Resolution Convolutional Neural Networks (SRCNN) [10], which train a 3 layer deep fully convolutional network end-to-end to achieve excellent SR performance. Kim *et al.* [17] used a very deep CNN network (20 weight layers) to achieve better performance and visual effects. In particular, they showed skip-connection and recursive convolution alleviate the burden of carrying identity information in the super-resolution network. In [18], Lim *et al.* propose the Enhanced Deep Residual Networks (EDSR) with better performance than SRResNet. Johnson *et al.* [19] proposed perceptual loss functions based on high-level features extracted from pretrained networks, which can reconstruct finer details compared to the per-pixel loss. Recently, Generative adversarial networks [20] have shown excellent results in many computer vision problems including SISR. Ledig *et al.* [12] used GAN to get photo-realistic natural images, which have better visually implausible performance than any other state-of-the-art methods. The authors propose a perceptual loss function constructed by both an adversarial loss and a perceptual content loss based on high-level features

 ISBN: 978-1-61208-716-0

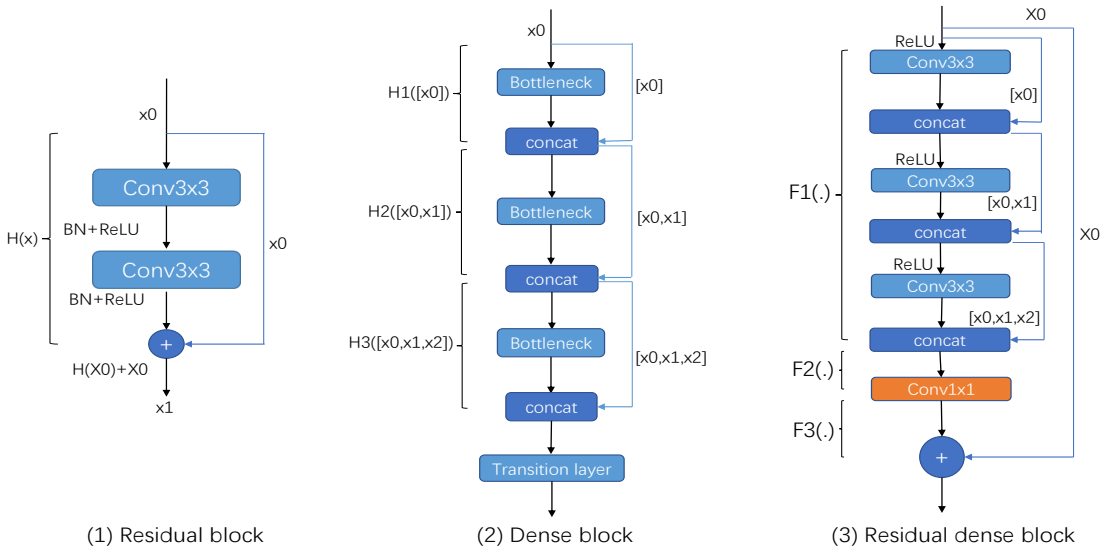

Figure 1. Examples of residual block, dense block and residual dense block. ”+” means element-wise summation operation. ”concat” means concatenation operation. ”bottleneck” in dense block which produces k feature-maps.

extracted from pre-trained Visual Geometry Group (VGG) networks.

III. METHOD

A. Network selection

Many research works [21][22] show networks that perform satisfactory in image generation, classification, and feature extraction and they are equally superior in image super-resolution. Among them, GANs, ResNets and DenseNets are successfully applied to image super-resolution tasks [23][24].

GANs: Following Goodfellow *et al.*, we define a discriminator network D_{θ_D} that is optimized in an alternating manner with the generator network G_{θ_G} to solve the adversarial minimum-maximum problem:

$$\min_{\theta_G}\max_{\theta_D}\mathbb{E}_{I^{HR}\sim p_{train}(I^{HR})}[logD_{\theta_D}(I^{HR})]+ \mathbb{E}_{I^{LR}\sim p_G(I^{LR})}[log(1-D_{\theta_D}(G_{\theta_G}(I^{LR}))] \quad (1)$$

where θ_G and θ_D represent the parameters of the generator and the discriminator, I^{HR} and I^{LR} represent the ground truth image and the low resolution image. The general idea is that it allows people to train a generative model G, the purpose of which is to fool the discriminator D that can distinguish between the real image and the generated image. With this approach, our generator can learn to create solutions that are highly similar to real images. This encourages perceptually superior solutions residing in the subspace, the manifold, of natural images.

ResNets: The main idea is to use a residual learning framework to ease the training of very deep networks. Let a single image x_0 go through a L-layer convolutional network. Each layer corresponds to a non-linear transformation $H_\ell(\cdot)$, where ℓ represents the index of the layer. Let x_ℓ be the output of ℓ-th layer. The traditional convolutional network generally uses the output of the ℓth layer as the input of the (ℓ+1)-th layer, which can be expressed as: $x_{\ell+1} = H_{\ell+1}(x_\ell)$. Unlike traditional CNNs, ResNets implements a residual block that sums up the identity mapping of the input to output of a layer, where the output can be depicted as: $x_{\ell+1} = H_{\ell+1}(x_\ell) + x_\ell$. This process eases the convergence during training. The structure of ResNets is shown in Figure 1(1).

DenseNets: The obvious difference between DenseNets and ResNets is that ResNets is a summation, while DenseNets is a concatenation. DenseNets enhances the transmission efficiency of information and gradients in the network. Each layer can directly get the gradient from the loss function and directly get the input signal, so that it can train deeper networks. The dense connection was introduced among memory blocks and dense blocks. Consequently, the feature maps of all previous layers are treated as separate inputs by connecting them to a single tensor $[x_0, x_1, ..., x_\ell]$, while their own feature maps are passed as input to all subsequent layers. Layer ℓ+1 receives the feature maps of all previous layers and can be expressed as: $x_{\ell+1} = H_{\ell+1}([x_0, x_1, ..., x_\ell])$. Figure 1(2) shows an example of dense block construction.

RDBs mainly integrates the residual blocks and the dense blocks. The structure difference is obvious in Figure 1(3). Let F_{l-1} and F_l represent the input and output of the ℓ-th RDB, respectively, and they all have G0 feature maps. In our experiment, we set G0 to 128. In the ℓ-th RDB, the output of the c-th convolutional layer can be formulated as:

$$F_{\ell,c} = \sigma(W_{\ell,c}[F_{\ell-1}, F_{\ell,1}, ..., F_{\ell,c-1}]) \quad (2)$$

where σ represents the Rectified Linear Unit (ReLU) activation function. $W_{\ell,c}$ is the weight of the c-th convolutional layer. For convenience, we ignore the bias term. In our work, we set the number of convolution layers in each RDB to 9. We use these layers to extract continuous memory. Then, connect all the feature maps extracted earlier. Inspired by MemNet [25], we introduce a 1×1 convolutional layer to adaptively control

 ISBN: 978-1-61208-716-0

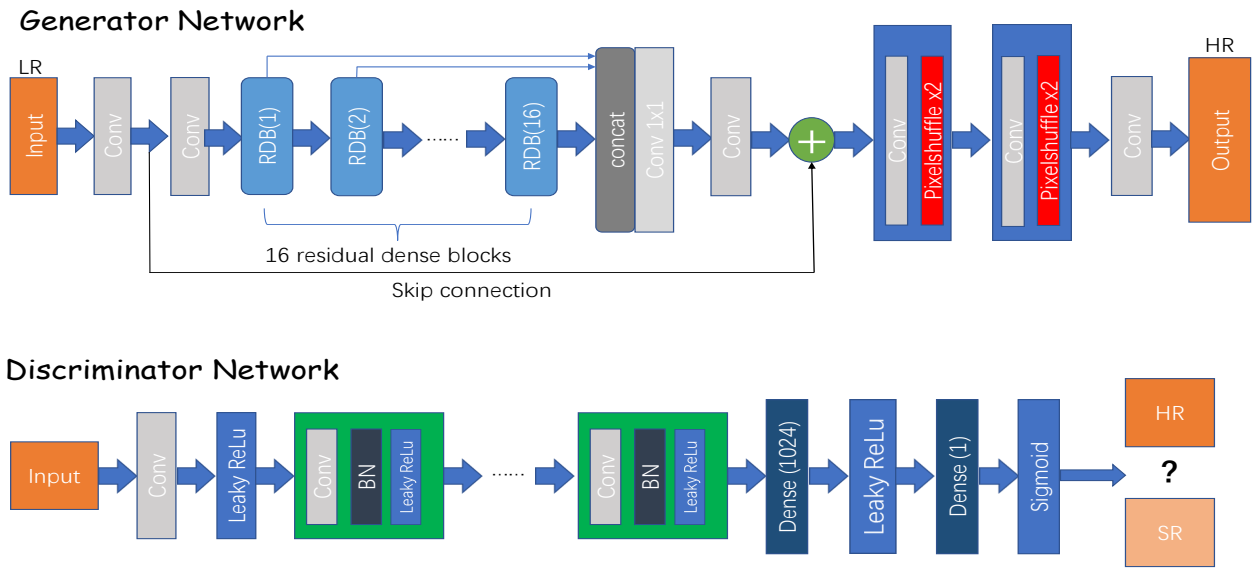

Figure 2. Architecture of Generator and Discriminator Network

the output information. Finally, the number of feature maps becomes G0. This step can be expressed as:

$$F_{\ell,m} = H_{LFF}([F_{\ell-1}, F_{\ell,1}, ..., F_{\ell,9}]) \quad (3)$$

where H_{LFF} represents a function of the 1*1 convolution layer. Finally, use the principle of residuals to achieve local residual learning. The output of the ℓ-th RDB can be expressed as:

$$F_\ell = F_{\ell-1} + F_{\ell,m} \quad (4)$$

B. Basic network architecture

The entire network structure of the generator is presented in Table I. We set up sixteen RDBs and each RDB is set as described above. In order to prevent the loss of LR images detail, we removed the pooling layer and the BN layer, then connected the outputs of all RDBs, with 1×1 convolution kernels to fuse feature maps and add residuals connect to retain more details. More details can be seen in Figure 2. This model can accept the input of LR images of any size, obtain the SR image of a given scaling factor α through the whole generator, and upgrade the image quality through continuous optimization of the generator and discriminator.

C. Loss function

The loss function we use is the same as Ledig *et al.* [12], combining pixel-wise loss and vgg19 loss [19] based on the high-level features extracted from the pre-trained 19 layer VGG networks. Given the high resolution ground truth image I^{HR}, the corresponding low resolution image I^{LR} and the image I^{SR} generated by our network, the loss function can be defined as follows:

$$L_{percep}(I^{HR}, I^{SR}) = \lambda_M \times L_M(I^{HR}, I^{SR}) + \lambda_V \times L_V(I^{HR}, I^{SR}) \quad (5)$$

where $L_M(I^{HR}, I^{SR})$ is pixel-wise loss and $L_V(I^{HR}, I^{SR})$ is vgg19 loss. λ_M and λ_V are scaling hyperparameters. In our work, we set λ_M to 1 and λ_V to 0.006.

TABLE I. ARCHITECTURE DETAILS FOR 4× RDGAN GENERATOR. NOTE THAT EACH "CONV" LAYER SHOWN IN THE TABLE CORRESPONDS THE SEQUENCE RELU-CONV.

Layers	Output size	Residual DenseNet	Feature-maps
Convolution	W×H	3×3 conv	64
RDB(1-16)	W×H	Bottleneck × 9	64
Concat	W×H	Connect	1024(64×16)
Convolution	W×H	1×1 conv	64
Convolution	W×H	3×3 conv	64
Summation	W×H	3×3 conv	64
Convolution	W×H	3×3 conv	256
Upscale	2W×2H	PixelShuffle	64
Convolution	2W×2H	3×3 conv	256
Upscale	4W×4H	PixelShuffle	64
Convolution	4W×4H	3×3 conv	3

pixel-wise loss: It is the Euclidean distance between the generated image I^{SR} and the ground truth image I^{HR}. Pixel-wise loss is defined as follows:

$$L_M(I^{HR}, I^{SR}) = \frac{1}{SWH}\|I^{HR} - I^{SR}\|^2 \quad (6)$$

where SWH is the size of the target image. This loss is added to achieve smoother textures from the ground truth image.

vgg19 loss: It is the Euclidean distance between the feature maps generated by the loss network. When given the pre-training network ϕ and a series of convolutional layers C and the feature map of each convolutional layer on C is $S_i \times W_i \times H_i$, we can define vgg19 loss as follows:

$$L_V(I^{HR}, I^{SR}) = \sum_{i\in C} \frac{1}{S_i \times W_i \times H_i}\|\phi_i(HR) - \phi_i(SR)\| \quad (7)$$

where $S_i \times W_i \times H_i$ represent the size of the respective feature map in the VGG networks.

adversarial loss: In addition to the perceptual loss described above, we also add the adversarial loss to the perceptual loss. This encourages our network to preserve more textures on

 ISBN: 978-1-61208-716-0

TABLE II. QUANTIFIED PERFORMANCE OF DIFFERENT SUPER-RESOLVED METHODS ON BENCHMARK DATA, WHICH IS MEASURED BY (PSNR [DB], SSIM). [4× UPSCALING].

Set5	Bicubic	Aplus [26]	SRCNN [10]	VDSR [17]	DRCN [27]	SRGAN [12]	**RDGAN**
PSNR	28.42	30.28	30.07	31.35	31.53	29.40	**31.70**
SSIM	0.8104	0.8603	0.8627	0.8838	0.8854	0.8472	**0.8903**
Set14							
PSNR	25.99	27.32	27.18	28.01	28.02	26.02	**28.13**
SSIM	0.7027	0.7491	0.7503	0.7674	0.7670	0.7397	**0.7872**
BSD100							
PSNR	25.96	26.82	26.68	27.27	27.23	25.18	**27.39**
SSIM	0.6675	0.7087	0.7101	0.7251	0.7233	0.6688	**0.7290**
Urban100							
PSNR	23.14	24.32	24.52	25.18	25.14	-	**25.68**
SSIM	0.6577	0.7183	0.7221	0.7524	0.7510	-	**0.7712**

natural images. It optimizes parameters by minimizing the generative loss L_{GAN} defined based on $D_{\theta_D}(G_{\theta_G}(I^{LR}))$, which means the probability of the discriminator that the reconstructed images $G_{\theta_G}(I^{LR})$ is a natural HR image:

$$L_{GAN} = \sum_{n=1}^{N} -logD_{\theta_D}(G_{\theta_G}(I^{LR})) \tag{8}$$

Finally, our loss function can be expressed as:

$$\begin{aligned} L(I^{HR}, I^{SR}) &= L_{percep}(I^{HR}, I^{SR}) + \\ &\quad \lambda_{GAN} \times L_{GAN}(I^{HR}, I^{SR}) \end{aligned} \tag{9}$$

We set λ_{GAN} = 0.001. After doing this, we get the images with more natural textures and more realistic details.

IV. EXPERIMENTS

A. Training Details

The train and validation datasets were sampled from DIV2K datasets [28]. DIV2K datasets were obtained from [29][30]. The train dataset has 800 images and the validation dataset has 100 images. We obtained the LR images by downsampling the HR images using bicubic kernel with downsampling factor r=4. This corresponds to a 16× reduction in image pixels. We test the performance on four standard benchmark datasets: Set5 [31], Set14 [32], BSD100 [33], Urban100 [34].

All the experiments were implemented by means of Python 3.6 and PyTorch [35] on a NVIDIA 1080Ti GPU. For training, we use the Red-Green-Blue (RGB) input patches of size 128×128 from LR images with the corresponding HR patches. Note that we can apply the generator model to images of arbitrary size as it is fully convolutional. We train our model with the ADAM optimizer [36] by setting β_1=0.9. The learning rate was initially set to 0.0001 and decreased by a factor of 10 after 50 epoches. We alternate updates to the generator and discriminator network, which is equivalent to k=1 as used in Goodfellow *et al.* [20]. A mini-batch size of 5 was set during the training. It takes about one days to train RDGAN.

B. Evaluation on benchmark datasets

We train all models with 400 epochs. The training process stopped after no improvements of the loss was observed after 350 epoches. We present the quantitative evaluation results of our RDGAN on public benchmark datasets in Table II. We compare the proposed method with the state-of-the-art methods including Aplus [26], SRCNN [11], SRGAN [12], VDSR [17] and DRCN [27]. For comparison, the SR results are evaluated with PSNR and Structural Similarity (SSIM) [37] on Y channel (i.e., luminance) of transformed YCbCr space. Our RDGAN shows significant improvement compared to other models. We also provide the qualitative results in Figure 3. We can see that the method we propose produces relatively sharper edges, while other models may produce ambiguous results.

V. CONCLUSION

In this work, we proposed a very deep Residual Dense Generative Adversarial Network (RDGAN) for Single Image Super-Resolution, where RDBs are used as basic modules for the generator network. By using the new generator network architecture, we maintain the accuracy of the reconstructed image while maintaining the visual quality of the super-resolution image. In terms of the loss function, we retain the confrontation loss, which makes the generated image retain full detail and more realistic in terms of visual perception. We evaluated our method on a large number of datasets and the results show that our RDGAN can achieve good results in Single Image Super-Resolution.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (NSFC), grant (NO. 11571044, 11471052, 11671052). Tianping Shuai is the corresponding author of this paper.

REFERENCES

[1] D. G. S. B. M. Irani, "Super-resolution from a single image," in Proceedings of the IEEE International Conference on Computer Vision, Kyoto, Japan, 2009, pp. 349–356.

[2] W. Shi et al., "Cardiac image super-resolution with global correspondence using multi-atlas patchmatch," in International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2013, pp. 9–16.

[3] T. M. Lehmann, C. Gonner, and K. Spitzer, "Survey: Interpolation methods in medical image processing," IEEE transactions on medical imaging, vol. 18, no. 11, 1999, pp. 1049–1075.

[4] F. Juefei-Xu and M. Savvides, "Single face image super-resolution via solo dictionary learning," in 2015 IEEE International Conference on Image Processing (ICIP). IEEE, 2015, pp. 2239–2243.

[5] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," arXiv preprint arXiv:1710.10196, 2017.

 ISBN: 978-1-61208-716-0

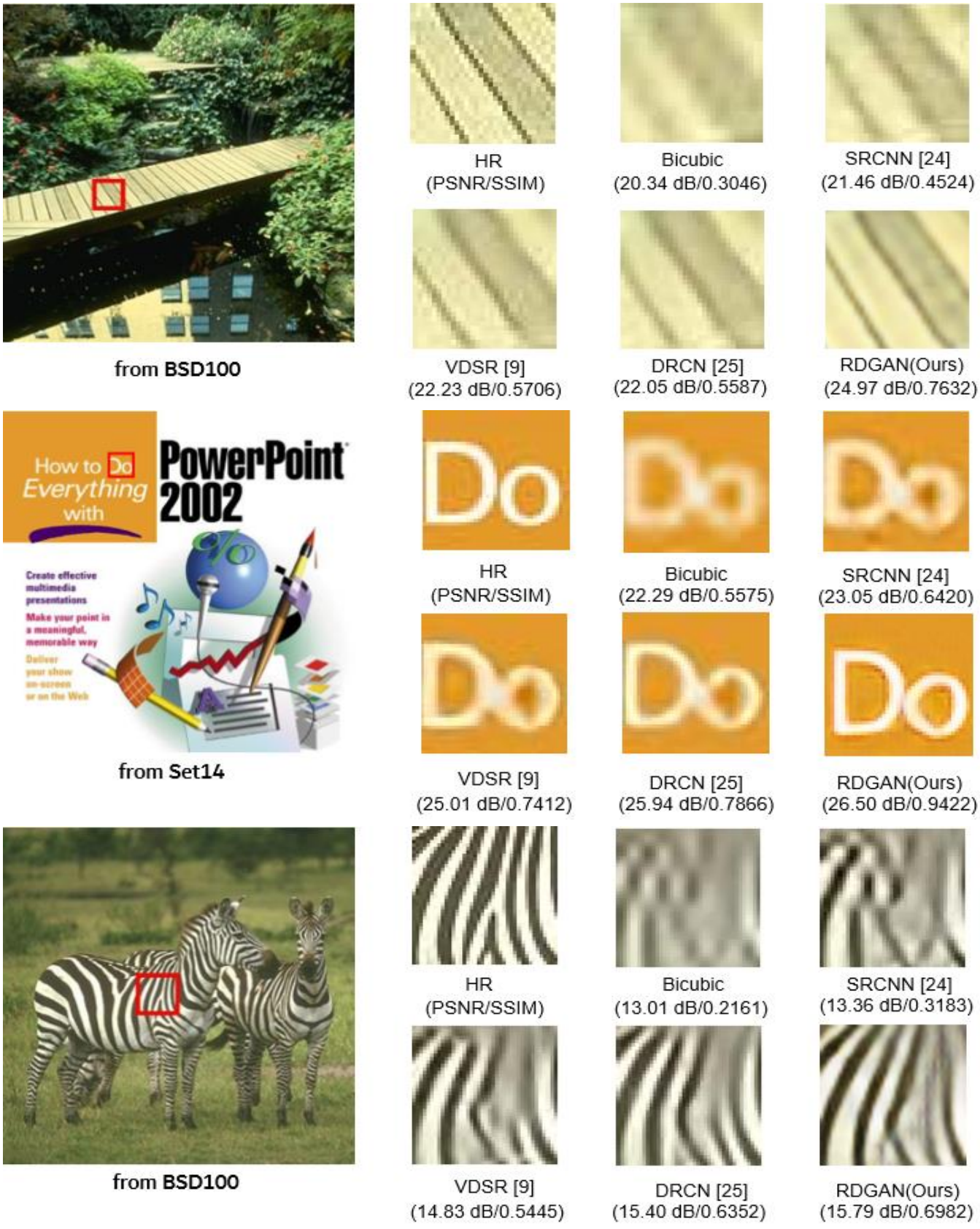

Figure 3. Qualitative comparison of our RDGAN with other methods on 4 super-resolution

 ISBN: 978-1-61208-716-0

[6] R. Keys, “Cubic convolution interpolation for digital image processing,” IEEE transactions on acoustics, speech, and signal processing, vol. 29, no. 6, 1981, pp. 1153–1160.

[7] H. A. Aly and E. Dubois, “Image up-sampling using total-variation regularization with a new observation model,” IEEE Transactions on Image Processing, vol. 14, no. 10, 2005, pp. 1647–1659.

[8] S. Schulter, C. Leistner, and H. Bischof, “Fast and accurate image upscaling with super-resolution forests,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3791–3799.

[9] T. Tong, G. Li, X. Liu, and Q. Gao, “Image super-resolution using dense skip connections,” in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 4799–4807.

[10] C. Dong, C. C. Loy, K. He, and X. Tang, “Learning a deep convolutional network for image super-resolution,” in European conference on computer vision. Springer, 2014, pp. 184–199.

[11] ——, “Image super-resolution using deep convolutional networks,” IEEE transactions on pattern analysis and machine intelligence, vol. 38, no. 2, 2016, pp. 295–307.

[12] C. Ledig et al., “Photo-realistic single image super-resolution using a generative adversarial network,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4681–4690.

[13] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, “Residual dense network for image super-resolution,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2472–2481.

[14] R. Keys, “Cubic convolution interpolation for digital image processing,” IEEE transactions on acoustics, speech, and signal processing, vol. 29, no. 6, 1981, pp. 1153–1160.

[15] H. Chang, D.-Y. Yeung, and Y. Xiong, “Super-resolution through neighbor embedding,” in Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004., vol. 1. IEEE, 2004, pp. I–I.

[16] K. Hayat, “Multimedia super-resolution via deep learning: A survey,” Digital Signal Processing, 2018.

[17] J. Kim, J. Kwon Lee, and K. Mu Lee, “Accurate image super-resolution using very deep convolutional networks,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 1646–1654.

[18] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, “Enhanced deep residual networks for single image super-resolution,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 136–144.

[19] J. Johnson, A. Alahi, and L. Fei-Fei, “Perceptual losses for real-time style transfer and super-resolution,” in European Conference on Computer Vision. Springer, 2016, pp. 694–711.

[20] I. Goodfellow et al., “Generative adversarial nets,” in Advances in neural information processing systems, 2014, pp. 2672–2680.

[21] J. Kim, J. K. Lee, and K. M. Lee, “Deeply-recursive convolutional network for image super-resolution,” 2015, pp. 1637–1645.

[22] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4700–4708.

[23] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” in Advances in neural information processing systems, 2015, pp. 91–99.

[24] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs,” IEEE transactions on pattern analysis and machine intelligence, vol. 40, no. 4, 2018, pp. 834–848.

[25] Y. Tai, J. Yang, X. Liu, and C. Xu, “Memnet: A persistent memory network for image restoration,” in Proceedings of the IEEE international conference on computer vision, 2017, pp. 4539–4547.

[26] R. Timofte, V. De Smet, and L. Van Gool, “A+: Adjusted anchored neighborhood regression for fast super-resolution,” in Asian conference on computer vision. Springer, 2014, pp. 111–126.

[27] J. Kim, J. Kwon Lee, and K. Mu Lee, “Deeply-recursive convolutional network for image super-resolution,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 1637–1645.

[28] E. Agustsson and R. Timofte, “Ntire 2017 challenge on single image super-resolution: Dataset and study,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 126–135.

[29] A. Ignatov, R. Timofte et al., “Pirm challenge on perceptual image enhancement on smartphones: report,” in European Conference on Computer Vision (ECCV) Workshops, January 2019.

[30] E. Agustsson and R. Timofte, “Ntire 2017 challenge on single image super-resolution: Dataset and study,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 126–135.

[31] M. Bevilacqua, A. Roumy, C. Guillemot, and M. L. Alberi-Morel, “Low-complexity single-image super-resolution based on nonnegative neighbor embedding,” 2012.

[32] R. Zeyde, M. Elad, and M. Protter, “On single image scale-up using sparse-representations,” in International conference on curves and surfaces. Springer, 2010, pp. 711–730.

[33] D. Martin et al., “A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics.” Iccv Vancouver:, 2001.

[34] J.-B. Huang, A. Singh, and N. Ahuja, “Single image super-resolution from transformed self-exemplars,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 5197–5206.

[35] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, “Automatic differentiation in pytorch,” 2017.

[36] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” arXiv preprint arXiv:1412.6980, 2014.

[37] Z. Wang et al., “Image quality assessment: from error visibility to structural similarity,” IEEE transactions on image processing, vol. 13, no. 4, 2004, pp. 600–612.

 ISBN: 978-1-61208-716-0

Concurrent Real-Time Object Detection on Multiple Live Streams Using Optimization CPU and GPU Resources in YOLOv3

Samira Karimi Mansoub
Mavinci Informatics Inc.
Ankara, Turkey
e-mail: samira.karimi@mavinci.com.tr

Rahem Abri
Mavinci Informatics Inc.
Ankara, Turkey
e-mail: rahem.abri@mavinci.com.tr

Anıl Hakan Yarıcı
Mavinci Informatics Inc.
Ankara, Turkey
e-mail: anil.yarici@mavinci.com.tr

***Abstract*— Recently, You Look Only Once, version 3 (YOLOv3) approach has been presented as a more efficient solution in the process of object detection. Despite the fact that YOLOv3 can obtain faster and more accurate results than other approaches, it needs to be used in a system with a single powerful Graphics Processing Unit (GPU). However, sometimes, there is a need to process multiple real-time object detection algorithms concurrently on a single GPU, where each object detection algorithm receives a live stream from a camera. It is challenging to have concurrent object detection from live streams on a single GPU. In this paper, we propose a two-step solution to this problem. In the first step, our goal is to provide a model to optimize memory usage and, in the second step, we propose a multi-thread approach that uses YOLOv3 to perform real-time object detection on multiple, concurrent, live streams on a single GPU. In this approach, GPU resources are optimally used. The proposed approach is evaluated on a public dataset and the result shows improvements in performance by an average of 12% in Central Processing Unit (CPU) usage and 13% in frames per second (FPS) compared to the YOLOv3.**

Keywords-Real-Time Object Detection; YOLO; Multi-Thread Approaches.

I. INTRODUCTION

Recent researches in the field of object detection based on Convolutional Neural Network (CNN) methods such as Region-CNN (R-CNN) [1], Faster R-CNN [3], YOLO [2], YOLOv2 [4] and YOLOv3 [5] have shown improvements when compared with other detection methods which focus on traditional detection [8][11]-[13]. Generally in object detection methods, the following steps are performed: 1) feature extraction of images [7][9], 2) classification [1][14][15] and 3) localization [6][10]. Recent approaches, such as YOLOv3, detect objects using the conventional system with a single GPU. YOLOv3 uses YOLOv2 as a base structure along with 53 convolutional layers. YOLOv3 is more powerful and faster than YOLOv2 because it uses the GPU in a more efficient way.

Although the YOLOv3 is more accurate than other approaches, it needs to be executed in a system with a single powerful GPU. In addition, in some applications, it is desired to use CNN's potential for concurrent, real-time object detection. Hence, an efficient hardware implementation along with an efficient network design are required. It is challenging to leverage this approach for concurrent, real-time object detection. For example, in some real-world applications, such as concurrent real-time inference on a GPU server in a commercial system, the available computing GPU resources are limited in terms of memory. In this case, each object detection approach receives a live stream from a camera and all processing is performed on a GPU. The main problem is system resources such as memory, CPU, and GPU when using them concurrently for real-time detection scenarios.

Alternative methods suggest using multiple GPUs in parallel. To solve this problem, in this paper, a model (network architecture design) is proposed, which uses YOLOv3 for concurrent real-time objection detection on a single GPU server using a multi-thread architecture. Our purpose is to provide an optimized architecture that significantly decreases memory usage while at the same time increasing the number of object detection outputs. Therefore, this architecture makes possible use of multiple YOLOv3s on a single GPU server with low memory usage and high speed in detection performance. We have implemented the proposed model on the Nvidia Quadro p5000 with Linux customized by Nvidia by the Compute Unified Device Architecture (CUDA) architecture. The result shows improvements in performance by an average of 12% in CPU usage and 13% in FPS compared to the YOLOv3.

The paper is organized as follows. In Section 2, related works are discussed. Problem description and methodology are presented in detail in Section 3. In Section 4, the experimental result is presented. The conclusion is drawn in Section 5.

II. RELATED WORK

Convolutional neural networks offer significant improvements for applications such as image classification, object detection [30], face detection [24], segmentation [31] and object tracking [32][19]. Traditional methods for object detection have focused on Scale-Invariant Feature Transform (SIFT) such as the Fast Point Feature Histograms (FPFH) [20] and Normal Aligned Radial Features (NARF) [21] that are used in 3D image registration. Classification methods are also used for object detection and include nearest-neighbor methods [22] and support vector machines [23].

 ISBN: 978-1-61208-716-0

More recent methods using convolutional neural networks [29] such as YOLO [2], YOLOv2 [4] and YOLOv3 [5] object detection approaches have shown more improvement in comparison with traditional computer vision methods such as methods based on SIFT [16]. Using deep learning-based strategies [33], these methods provide significant speed advantages over R-CNN (for example, 45 frames per second in YOLO on an Nvidia Titan-X GPU). Although convolutional neural networks require a training process, they can be applied for general challenges. They can also be used along with fewer hardware resources.

Despite an increase in speed, there is a gap between software and hardware implementations [34] due to high power consumption. In fact, there is a need to implement hardware along with an efficient neural network design in order to exploit CNN's for low-power. Some research exists on real-time processing using multi-core architectures and GPUs. Methods such as the Gaussians Mixture Model (GMM) for background modeling are used in [17] based on such architectures. In the GPU architecture, Nvidia has provided GeForce, Quadro and Tesla/Fermi series with different performance ranges. In [25], a hardware architecture for real-time object detection using depth and edge information is proposed. In [26], an analytical framework (OPTiC) is proposed for partitioning optimal CPU-GPU co-execution on systems. In [27], a GPU-based floating real-time object detection system is proposed. Based on the results in this paper, using GPU instead of CPU can increase speed and improve performance. Using convolutional neural networks in other applications like Internet of Things (IoT) and mobile edge computing is discussed in [28]. In this paper, real-time multiple object tracking is implemented on an Nvidia GPU architecture.

There has been a very limited number of works on concurrent real-time object detection on multiple live streams. In this paper, we try to provide an optimized architecture to concurrently detect objects on multiple cameras. In this architecture, YOLOv3 is used as a backbone approach to detect objects. We conducted experiments to evaluate the approach. In the next section, we describe the methodology.

III. Problem Description and Methodology

In this section, we discuss two improvements in the performance of the YOLO in regards to the process of object detection. In the first subsection, we present the new architecture model which changes the structure of YOLO in the object detection process to reduce the number of unnecessary computations and conversions in the object detection process. In the second subsection, we provide a multi-thread approach for performing real-time object detection on multiple live streams concurrently. Using these two improvements, we achieve increased efficiency in the object detection process using YOLOv3.

A. The Architecture of the Model

In this subsection, we plan to explain the network architecture of the object detection process. Before we start to discuss the model, we need to describe the architecture of the object detection process in YOLOv3. Then, we discuss the problems of the architecture. Finally, we provide a model as a solution to overcome these problems.

The main architecture of the YOLOv3 is depicted in Figure 1. This architecture includes an object detection process using YOLO on the darknet framework. As shown in Figure 1, there are computing and converting processes. All of the computing processes are done on the GPU side and the converting processes are conducted on the CPU side. The process is started with converting RGB (red, green, and blue) image as a three dimensions matrix to the YOLO image using a function. The YOLO image is a two dimensions matrix that contains RGB values for each pixel. The main function of the object detection is done using YOLO on a converted YOLO image. The output of the detection is a vector of coordinates for each detected object. As previously mentioned, YOLOv3 uses a conventional neural network to detect objects. Since in this paper our goal is to discuss the architecture of the object detection process, we avoid the explanation of the YOLOv3 method in detail. Finally, the CPU converts the YOLO image to RGB image format.

Although the YOLOv3 can obtain more accuracy and speed compared to other approaches, it needs to be used in a system with a powerful single GPU. Also, there are unnecessary computations and conversions in the object detection process in the Darknet framework. Here, we provide an improved model to optimize the process of the detection using the YOLO in the Darknet framework. We call the improved model MvcYOLO. The MvcYOLO model is presented in Figure 3. In this model, we change the detected coordinates using YOLO detection algorithm and eliminate the unnecessary conversion at the last step of the Darknet framework. For each detected object, there is a coordinate point which describes the center of each object and two scales as the percentage of the specified object's length and width. These coordinates are produced with YOLO. In this step, we intervene in the normal process of YOLO and calculate new coordinates instead of drawing detected regions by using YOLO provided coordinates. For this propose, we calculate the left corner, length, and width of the detected region using the information of provided coordinates by YOLO. As mentioned before, in YOLO detection, we have the points that belong to the center of each detected object and two percentage of length and width in every detected object. In this process, we convert YOLO detection coordinates to our desired coordinates. Our purpose is to draw regions obtained from detected objects on the RGB image directly. In fact, YOLO makes the additional computations with drawing detected regions on the YOLO image and then unnecessary converts the YOLO image to RGB image. So, to achieve more efficiency, we prevent drawing regions by YOLO on YOLO image and re-draw regions using the new estimated coordinates directly

 ISBN: 978-1-61208-716-0

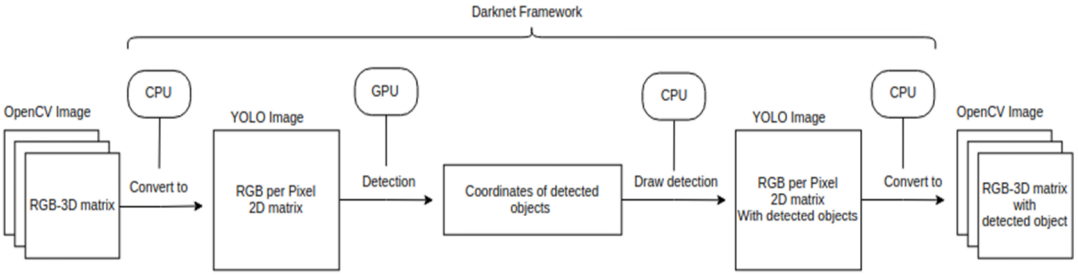

Figure 1. The architecture of the object detection process in YOLOv3.

on the OpenCV RGB image. Finally, we consider new coordinates to draw regions on RGB image directly using the estimated coordinates. These extra calculations and conversions may be considered little, but we obtain improvements in terms of CPU usage by eliminating them. The results are discussed in the next section.

B. *Multi-thread Approach to Handle Concurrent Surveillance Videos*

The second part of our improvements is focused on the GPU usage per YOLO object. Each YOLOv3 object in the Darknet framework uses 1.70 GB GPU memory. This means that we can handle a surveillance video with 1.70 GB GPU memory. Because the GPU memory is limited in the GPU cards, it is a costly process to cover a large scale of surveillance cameras. This paper provides a multi-thread approach to handle concurrent surveillance videos with each YOLOv3 object. As mentioned before, if each surveillance camera works with a YOLOv3 object, 1.70 GB GPU of memory is needed to be assigned to it. As a result, it requires a lot of GPU memory for performing real-time object detection on multiple live streams concurrently.

To solve this problem, in this paper we present a multi-thread approach as a solution to handle concurrent surveillance videos. In this approach, a pre-processing and a post-processing step are added to the YOLO object. In the pre-processing step, a multiplexer for fragmentation of live streams is applied. In the post-processing also, a de-multiplexer is used to re-fragmentation and assembling the frames of each live stream. An overview of the approach is depicted in Figure 2.

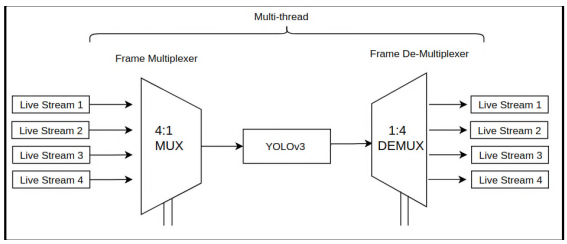

Figure 2. An overview of the multi-thread approach.

As shown in Figure 2, the input and output of YOLO are defined as Y_{In} and Y_{out}. There are a multiplexer and a de-multiplexer at the beginning and at the end of the process. In the multiplexer, in each live stream, a frame of every four frames is selected. Thus, there are intervals of four frames in each live stream. In the next step, the frames are passed to detect objects using YOLOv3. In this way, frames of four live streams are fragmented and are passed separately to the YOLOv3 algorithm to detect the objects. The output of the YOLOv3 algorithm is a video that contains every four video frames. This fragmentation process works as a multi-thread approach. To investigate the efficiency, we evaluated the approach using experiments and the results are presented in the next section.

IV. EVALUATION METHODOLOGY AND EXPERIMENTS

In this section, we evaluate the methodology and experiments in two subsections. The first subsection discusses the experiment environment. The second subsection presents experiments to evaluate the proposed model and approach in Section III by describing an architecture structure.

A. *Experiment Environment*

In this section, we evaluate the performance of the proposed model and architecture in the previous section. First, we prepare an experiment to compare the efficiency between the YOLOv3 and MvcYOLO. Our evaluation is on a dataset of 2D MOT 2015 containing 30 videos with FPS 30 and different qualities resolution (1280×720, 1920×1080 and 3840×2160 pixels). This benchmark contains video sequences in different environments. The properties of a video analysis server used as a test server are presented in Table I.

TABLE I. THE PROPERTIES OF VIDEO ANALYSIS SERVER.

CPU	Intel core i9-7940X 14 core/ 28 thread
GPU	Nvidia Quadro p5000 16GB
RAM	32 GB DDR4
Disk	256 NVM Express SSD

 ISBN: 978-1-61208-716-0

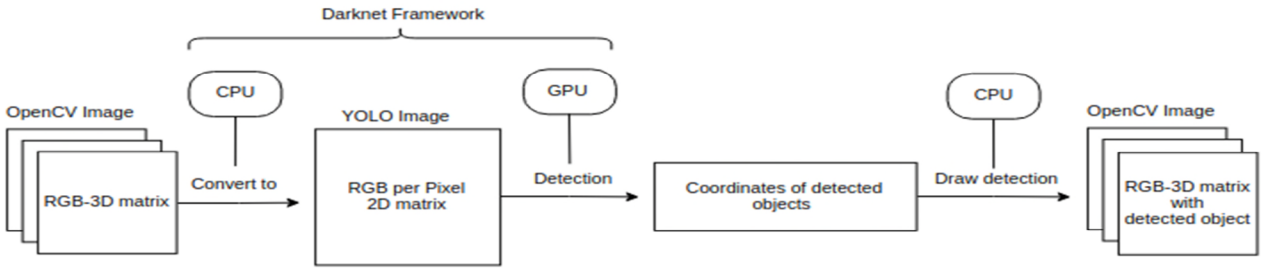

Figure 3. The architecture of the object detection process in MvcYOLO model.

B. Experimental Evaluation of the Model

As discussed before in this section, in order to evaluate the proposed model MvcYOLO, we make an experimental comparison between the original YOLOv3 and MvcYOLO. Our focus is on system resources like CPU, GPU, RAM and FPS values. CPU is one of the critical resources in this area because it performs the main conversion steps in the darknet framework. As previously mentioned, we prepare 10 videos in a different range of quality. The results are presented in Figure 4 in three-line graphs denoted as a, b, and c. The first line graph (a) compares the CPU usage percentage between original YOLOv3 and MvcYOLO in different video qualities. Overall, the CPU usage increases over the video qualities during the increasing computational power for handling large size of frames. The CPU usage in the MvcYOLO is lower than YOLOv3 overall video qualities. The bigger drop of CPU usage was seen on 4k video quality. We conclude that MvcYOLO has a better performance on high-quality videos.

The second line graph (b) uses the frame rate as a criterion for evaluation. The FPS is the frequency at which consecutive images, called frames, appear on a display. The FPS is compared in two models and the result shows the average of FPS values in the two approaches. It can be seen that the FPS of MvcYOLO was far higher than the original YOLOv3. The MvcYOLO improves FPS of high-quality video frames 13 to 26 in comparison to the original YOLOv3 approach. However, the MvcYOLO uses lower computational power in terms of CPU and RAM resources in each frame. As illustrated in the results, CPU usage has a high correlation with FPS. In 720p, the FPS value of YOLOv3 and MvcYOLO are 28, while this amount decreases steeply to 26 and continues to decline but more gradually to 13 in 4k video quality on YOLOv3 approach. In contrast, the FPS amounts of MvcYOLO declined more steeply to 26 in 1080p and 4k video qualities.

The third line graph (c) shows the average RAM usage in terms of GB over different video qualities in each approach. MvcYOLO uses less RAM than YOLOv3 because of doing less computation in the conversion process while improvements are not significant. Changes in RAM usage of YOLOv3 and MvcYOLO are approximately the same in each video quality.

C. Experimental Evaluation of the Multi-Thread Model

In the second part of the evaluation, our goal is to consider the proposed model in the previous section with a multi-thread approach as a solution to handle concurrent surveillance videos. As discussed in Section III, to obtain more efficiency in YOLO, we need to use a multi-thread architecture. So, we performed experiments to find the optimal number of cameras. Based on the result, 8GB GPU memory could handle a maximum of four surveillance cameras simultaneously. This paper provides a multi-thread approach to handle a minimum of four surveillance cameras over each YOLOv3 object and supports 16 cameras for 8GB GPU memory. The reason for choosing four cameras is the thread scheduler of the operating system and CPU power of the analysis server.

We call the multi-thread approach Multi-MvcYOLO. We compare MvcYOLO and Multi-MvcYOLO in three metrics of CPU usage, FPS and RAM usage. The results are depicted in Figure 5. Based on the results, it can be seen that CPU usage of Multi-MvcYOLO sharply decreased in each video quality especially on high-quality videos. In Multi-MvcYOLO the value of FPS is less compared to MvcYOLO because of performing four videos as multi-threads. However, Multi-MvcYOLO needs much more RAM than the others due to the thread scheduling process that uses extra RAM for scheduling.

As mentioned before, Multi-MvcYOLO handles four videos on a single YOLOv3 instance while YOLOv3 cannot handle more than one video on an object instance. Each YOLOv3 object instance uses approximately 1.70 GB from GPU memory. Since GPU memory is so costly, Multi-MvcYOLO approach uses optimal GPU memory and handles more surveillance cameras over a normal GPU card. Interestingly, Multi-MvcYOLO has better performance compared with YOLOv3 over the video qualities.

 ISBN: 978-1-61208-716-0

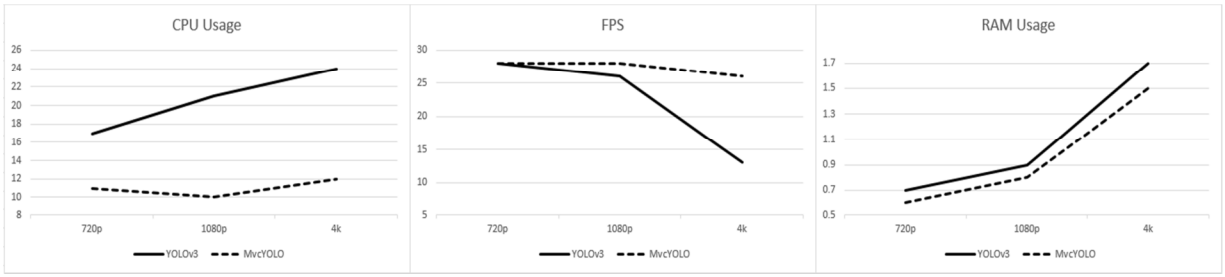

Figure 4. A comparison between original YOLOv3 and MvcYOLO in different video qualities on a) The CPU usage percentage b) FPS and c) RAM Usage

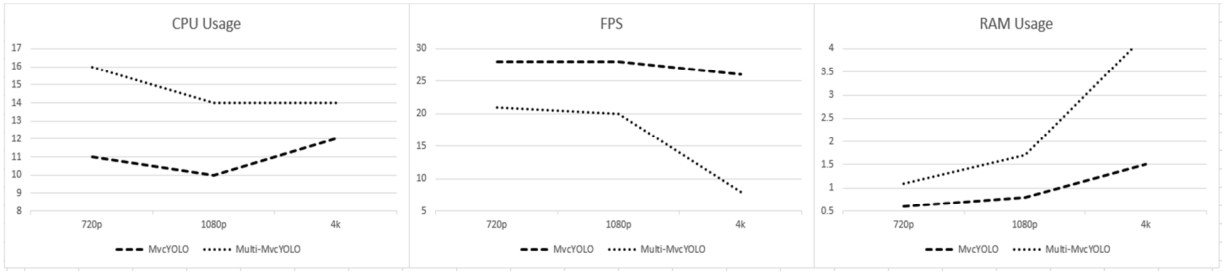

Figure 5. A comparison between original MvcYOLO and Multi-MvcYOLO in different video qualities on a) The CPU usage percentage b) FPS and c) RAM Usage

V. CONCLUSION AND FUTURE WORK

In the existing systems, it is necessary to use a powerful single GPU when applying YOLOv3 to detect objects. However, sometimes, there is a need to process multiple real-time object detection algorithms concurrently on a single GPU, where each object detection algorithm receives the live stream from a camera. In these cases, to reduce memory usage and other resources, we proposed solutions in two steps. In the first step, we provided a model for optimal memory usage and in the second step, we proposed a multi-thread approach that uses YOLOv3 to perform real-time object detection on multiple live streams concurrently on a GPU. This way, GPU resources are optimized to solve the limited memory issue. Experimental results show that the proposed approach can reduce memory consumption and increase performance by an average of 12% in CPU usage and 13% in FPS compared to the original YOLOv3. As future work, we are planning to improve MvcYOLO in terms of load balancing and network overhead on a large scale of camera surveillance environments.

ACKNOWLEDGMENT

This research is supported by Mavinci Informatics Inc. in Turkey. Mavinci is an R&D company working especially in information and communication technologies, security and defense areas with the capability of software development, artificial intelligence, and machine learning. The operational areas are; Intelligent Video Analytics based on Deep Learning, Nuclear Safety Research and Analysis, Disaster and Emergency Management, Decision Support Systems, Command and Control Systems, Chemical Biological Radiological and Nuclear Security Solutions, Image Processing and Project Management.

REFERENCES

[1] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” CVPR’14 Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp. 580-587, 2014.

[2] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol.1, pp. 779-788, 2016.

[3] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R- CNN: Towards real-time object detection with region proposal networks,” Advances in neural information processing systems (NIPS), vol. 39, pp. 1137-1149, 2015.

[4] J. Redmon, and A. Farhadi, “YOLO9000: better, faster, stronger,” Computer Vision and Pattern Recognition (CVPR), vol. 1, no. y, pp. 6517-6525, 2017.

[5] J. Redmon, and A. Farhadi, “YOLOv3: An Incremental Improvement,” Technical report, 2018.

[6] M. B. Blaschko, and C. H. Lampert, “learning to localize objects with structured output regression,” In Computer Vision ECCV, pp. 2-15, 2008.

[7] N. Dalal, and B. Triggs, “Histograms of oriented gradients for human detection,” In Computer Vision and Pattern

 ISBN: 978-1-61208-716-0

Recognition (CVPR), IEEE Computer Society Conference on, vol. 1, pp. 886-893. IEEE, 2005.

[8] T. Dean, M. Ruzon, M. Segal, J. Shlens, S. Vijayanarasimhan, and J. Yagnik, "Fast, accurate detection of 100,000 object classes on a single machine," In Computer Vision and Pattern Recognition (CVPR), IEEE Conference on, pp 1814-1821, 2013.

[9] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng et al, "Decaf: A deep convolutional activation feature for generic visual recognition,"vol. 32, pp. I-647-I-655, 2013.

[10] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," In International Conference on Learning Representations (ICLR2014), CBLS, abs/1312.6229, 2013.

[11] M. A. Sadeghi, and D. Forsyth, "30hz object detection with dpm v5," In Computer Vision ECCV 2014, pp. 65-79, Springer, 2014.

[12] J. Yan, Z. Lei, L. Wen, and S. Z. Li, "The fastest deformable part model for object detection," In Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on, pp. 2497-2504, IEEE, 2014.

[13] B. Benjdira, T. Khursheed, A. Koubaa, A. Ammar, K. Onuni, "Car Detection using Unmanned Aerial Vehicles: Comparison between Faster R-CNN and YOLOv3," Proceedings of the 1st International Conference on Unmanned Vehicle Systems (UVS), vol. 1, pp. 5386-9368, 2019.

[14] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part based models," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, pp. 1627-1645, 2010.

[15] P. Viola, and M. Jones, "Robust real-time object detection," International Journal of Computer Vision, vol. 4 pp. 34-47, 2001.

[16] D. Lowe, "Object recognition from local scale-invariant features," The Proceedings of the Seventh IEEE International Conference on Computer Vision, vol. 2, pp. 1150-1157, 1999.

[17] P. Kumar, A. Singhal, S. Mehta, and A. Mittal, "Real-time moving object detection algorithm on high-resolution videos using GPUs," Journal of Real-Time Image Processing, vol. 11, no.1, pp. 93-109, 2016.

[18] R.T. Collins, A. J. Lipton, T. Kanade, H. Fujiyoshi, D. Duggins, Y. Tsin et al, "VSAM: a system for video surveillance and monitoring," Technical Report CMU-RI-TR- pp. 00-12, Carnegie Mellon University, Pittsburgh, PA 2000.

[19] S. Veeraraghavan, and A. Chellappa, "Object detection, tracking and recognition for multiple smart cameras." Proc. IEEE, vol. 96, no. 10, pp. 1606-1624, 2008.

[20] R. Usu, R. B. Blodow, and M. Beetz, "Fast point feature histograms (FPFH) for 3D registration," In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA'09), IEEE Press, pp. 1848-1853, 2009.

[21] B. Steder, R. B. Rusu, K. Konolige, and W. Burgard, "NARF: 3D range image features for object recognition," In Proceedings of the Workshop on Defining and Solving Realistic Perception Problems in Personal Robotics at the IEEE/RSJ International Conference on Intelligent Robots and Systems, vol. 44, 2010.

[22] J. Liebelt, C. Schmid, and K. Schertler, "Viewpoint-independent object class detection using 3D feature maps," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'08), pp.1-8, 2008.

[23] M. R. Lata, and M. Y. Alvino, "FPGA implementation of support vector machines for 3D object identification," In Proceedings of the 19th International Conference on Artificial Neural Networks: Part I (ICANN'09), Lecture Notes in Computer Science, vol. 5768, Springer-Verlag, pp. 467-474. 2009.

[24] D. Han, J. Choi, J. Cho, and D. Kwak, "Design and VLSI implementation of high-performance face detection engine for mobile applications," In Proceedings of the IEEE International Conference on Consumer Electronics, pp. 705-706, 2011.

[25] K. Christos, T. Christos, and T. Theocharis, "A Hardware Architecture for Real-Time Object Detection Using Depth and Edge Information," ACM Transactions on Embedded Computing systems (TECS), vol. 13, no. 3, pp. 1-19, 2013.

[26] S. Wang, G. Ananthanarayanan, and T. Mitra, "OPTiC: Optimizing Collaborative CPU-GPU Computing on Mobile Devices with Thermal Constraints," IEEE transactions on computer-aided design of integrated circuits and systems, vol. 38 , no. 3 , pp. 393-406, 2018.

[27] Y. Jie, and M. Jian-min, "GPU Based Real-time Floating Object Detection System," 2nd International Conference on Electronics, Network and Computer Engineering 2016.

[28] B. Blanco-Filgueira, D. Garc´ıa-Lesta, M. Fernandez-Sanjurjo, and P. Lopez, "Deep Learning-Based Multiple Object Visual Tracking on Embedded System for IoT and Mobile Edge Computing Applications," ICDSC '18 Proceedings of the 12th International Conference on Distributed Smart Cameras, No. 22, 2018.

[29] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," Proceedings of the IEEE, vol. 86, no. 11, pp. 22782324, 1998.

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

[31] E. Shelhamer, J. Long, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 4, pp. 640-651, 2017.

[32] E. Gundogdu, and A. A. Alatan, "Good features to correlate for visual tracking," IEEE Transactions on Image Processing, vol. 27, no. 5, pp. 2526-2540, 2018.

[33] V. Sze, Y.-H. Chen, T.-J. Yang, and J. Emer, "Efficient Processing of Deep Neural Networks: A Tutorial and Survey," Proceedings of the IEEE, vol. 105, no. 12, pp. 2295-2329, 2017.

[34] S. W. Keckler, W. J. Dally, B. Khailany, M. Garland, and D. Glasco, "GPUs and the Future of Parallel Computing," IEEE Micro, vol. 31, no. 5, pp. 717, 2011.

 ISBN: 978-1-61208-716-0

Comparison of Corneal Pulse Entropy to Distinguish Healthy Eyes from Those with Primary Open-Angle Glaucoma

Michał M. Placek, Patryk M. Ząbkiewicz, and Monika E. Danielewska
Department of Biomedical Engineering, Faculty of Fundamental Problems of Technology,
Wrocław University of Science and Technology, Wrocław, Poland
Emails: michal.placek@pwr.edu.pl, 221394@student.pwr.edu.pl, monika.danielewska@pwr.edu.pl

***Abstract*—Corneal Pulsation (CP), as one of the manifestations of eye dynamics, has shown a great potential in the glaucoma diagnosis. The morphology of the CP signal, acquired noninvasively with a non-contact ultrasonic technique, has been found to alter in glaucoma patients. The aim of this preliminary study was to ascertain whether complexity of the CP signal may be a useful parameter to differentiate glaucoma patients from healthy individuals. *Refined Composite Multiscale Fuzzy Entropy* (RCMFE) was considered as a complexity measure. RCMFE of the CP signal was calculated in 25 glaucoma patients and 25 healthy subjects. Statistical analysis showed that, on average, glaucomatous eyes were characterized by higher entropy of the CP signal than healthy ones. This result suggests that RCMFE of the CP signal may support ophthalmologists in glaucoma diagnosis in its early stages.**

Keywords–multiscale entropy; corneal pulsation; glaucoma diagnosis.

I. INTRODUCTION

Corneal Pulsation (CP) reflects temporal corneal expansion related to ocular dynamics [1] and cardiovascular activity [2]. It has been shown that morphology of the noninvasively acquired CP signal changes with advancing age [3]. These specific changes, in the form of a double-peak wave in the CP signal during one heart cycle, have been named the Ocular Dicrotic Pulse (ODP) [4]. ODP has been detected more frequently in glaucoma patients than in healthy individuals, which has been explained by higher ocular rigidity of glaucomatous eyes [4][5].

Detection of ODP signal from the CP signal alone is not straightforward, since the CP signal is often noisy, irregular, and non-stationary. To overcome this problem, it has been proposed to synchronously measure reference cardiovascular signals—such as blood pulse or ECG, which, in contrast to the CP signal, exhibit clearly defined peaks—and to use, e.g., *Dynamic Time Warping* to synchronize these two classes of signals [3] and the wavelet transform to analyze the CP signal without the synchronization [6]. However, there is still a need to find such a feature of the CP signal, which would enable differentiating glaucoma patients from healthy subjects, without the necessity of explicit identification of the ODP signal or measurements of any auxiliary cardiovascular signal. To address this need, we propose to use the multiscale entropy algorithm to evaluate signal complexity.

Many previous studies have already shown that entropy of various physiological signals is altered in a wide range of pathological states [7][8]. To the best of our knowledge, entropy of the CP signal has never been considered so far. The aim of this preliminary study was to ascertain whether this entropy based approach applied to the CP signal is sufficient to differentiate Primary Open-Angle Glaucoma (POAG) patients from healthy individuals. The paper is organized as follows: Section 2 describes material and methods, Section 3 contains results, Section 4 discusses and concludes the study.

II. MATERIAL AND METHODS

The study sample consisted of 25 patients (8 males and 17 females) with POAG, aged from 59 to 79 (69.1 ± 6.0 years, mean ± SD), and a control, age-matched group of 25 healthy subjects (8 males and 17 females; 65.2 ± 6.4 years). Exclusion criteria were: systemic diseases, any previous ocular surgical procedure, conjunctival or intraocular inflammation, and corneal abnormalities such as edema or scars. Before the measurements, the purpose of the study and the protocol were explained to the participants. Afterwards, signed informed consent form was obtained from all patients and controls. The study had been approved by the Bioethics Committee of the Military Institute of Medicine in Warsaw (decision No. 67/WIM/2015) and adhered to the tenets of the Declaration of Helsinki.

Measurements of the CP signal of the glaucomatous eye in POAG patients and randomly selected eye in healthy subjects were performed using a non-contact ultrasonic distance sensor. The sensor allowed 10-second continuous data acquisition with the sampling frequency set to 400 Hz and *in-vivo* measurement of the CP amplitude with an accuracy below 1 μm [9]. For each eye, CP measurements were repeated five times and then three good-quality recordings, not affected by eye blink artifacts, as evaluated by an expert, were chosen for further analysis.

After preprocessing, including filtering in the range of 0.5–20 Hz and removing linear trend, signal entropy was estimated using the *Refined Composite Multiscale Fuzzy Entropy* ($RCMFE_\mu$) [8] up to scale factor No. 50. $RCMFE_\mu$ is defined as a sequence of fuzzy entropies computed for selected scale factors, where consecutive scales correspond to decreasing granularity of the analyzed time series. The average value of $RCMFE_\mu$ was calculated from three repeated measurements for each participant and scale factor. Since the null hypothesis of normality of $RCMFE_\mu$ was not rejected by the Kolmogorov–Smirnov test, the values obtained for each scale were compared between the two groups using the unpaired *t*-test. In the range of scales where *p*-values were below a significance level $\alpha = 0.05$, scale-averaged entropies

 ISBN: 978-1-61208-716-0

were computed for each participant and finally compared between the two groups using the unpaired *t*-test.

III. Results

Figure 1a shows group-averaged values of $RCMFE_{\mu}$ represented as a function of scale factor in the two groups, whereas Figure 1b presents results of statistical comparison between the groups for each scale factor. It can be seen that mean values of $RCMFE_{\mu}$ are significantly different between the groups for scales in the range of 25–34. Results of the unpaired *t*-test applied for entropies averaged over the abovementioned scale range are shown in Figure 2.

IV. Conclusion and Future Work

According to the decomplexification theory of illness, a loss in complexity is usually associated with pathological states [10]. Hence, it could be expected that a glaucomatous eye should be characterized by lower entropy of the CP signal than a healthy one. Our results, however, reveal the opposite tendency, which can be related to higher ocular stiffness (higher mechanical resistance of cornea) in glaucomatous eyes caused by the elevation of intraocular pressure [11]. Nevertheless, $RCMFE_{\mu}$ applied in this study satisfied our goal, as the measure was sensitive enough to differentiate between glaucoma patients and healthy subjects without the need for detecting the ocular dicrotism. To obtain results with higher degree of confidence, the analysis should be repeated on a larger study sample. More male participants need to be recruited if gender-specific differences are to be investigated. Ten-second measurements of the CP signal

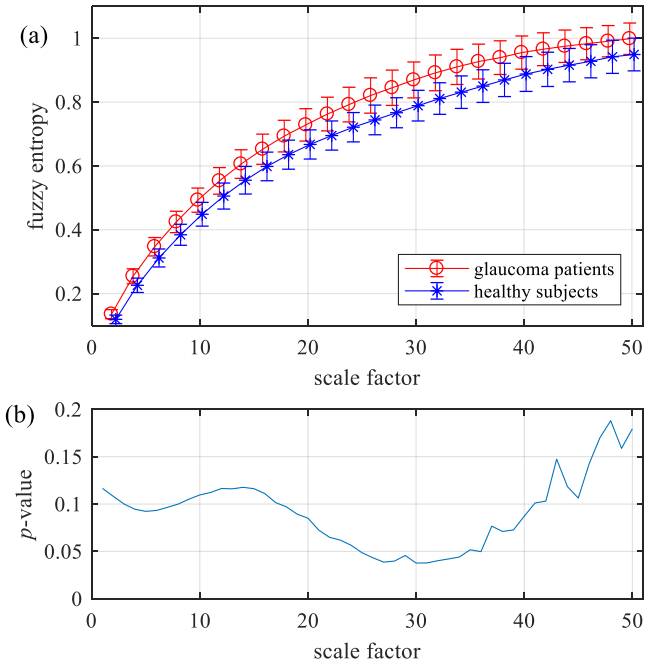

Figure 1. Refined composite multiscale fuzzy entropy ($RCMFE_{\mu}$) of the corneal pulse (CP) signal. (a) Means of grouped data ± 1.96 standard error of means. For better visibility, only even numbers of scale factor are shown. (b) *p*-values resulting from the comparison between the two groups using unpaired *t*-test.

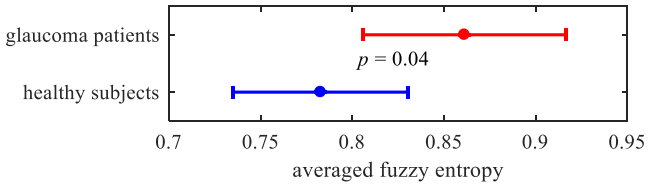

Figure 2. Mean values of scale-averaged entropies ± 1.96 standard error of means. Entropies were averaged in the scale range 25–34.

appear to be sufficiently long to reliably estimate entropy using $RCMFE_{\mu}$. The proposed $RCMFE_{\mu}$ of a CP signal may, in the future, constitute a sensitive indicator of changes in ocular stiffness and support ophthalmologists in glaucoma diagnosis in its early stages.

Acknowledgment

This research was supported by the National Centre for Research and Development, Poland (grant No. LIDER/074/L-6/14/NCBR/2015).

References

[1] M. A. Kowalska, H. T. Kasprzak, D. R. Iskander, M. E. Danielewska, and D. Mas, "Ultrasonic in vivo measurement of ocular surface expansion," *IEEE Transactions on Biomedical Engineering,* vol. 58, pp. 674–680, 2011.

[2] H. T. Kasprzak and D. R. Iskander, "Spectral characteristics of longitudinal corneal apex velocities and their relation to the cardiopulmonary system," *Eye,* vol. 21, pp. 1212–1219, 2007.

[3] M. E. Danielewska, D. R. Iskander, and P. Krzyzanowska-Berkowska, "Age-related changes in corneal pulsation: ocular dicrotism," *Optometry and Vision Science,* vol. 91, pp. 54–59, 2014.

[4] M. E. Danielewska, P. Krzyżanowska-Berkowska, and D. R. Iskander, "Glaucomatous and age-related changes in corneal pulsation shape. The ocular dicrotism," *PLOS ONE,* vol. 9, p. e102814, 2014.

[5] A. Hommer *et al.*, "Estimation of ocular rigidity based on measurement of pulse amplitude using pneumotonometry and fundus pulse using laser interferometry in glaucoma," *Investigative Ophthalmology & Visual Science,* vol. 49, pp. 4046–4050, 2008.

[6] T. Melcer, M. E. Danielewska, and D. R. Iskander, "Wavelet representation of the corneal pulse for detecting ocular dicrotism," *PLOS ONE,* vol. 10, p. e0124721, 2015.

[7] M. D. Costa and A. L. Goldberger, "Generalized multiscale entropy analysis: application to quantifying the complex volatility of human heartbeat time series," *Entropy,* vol. 17, pp. 1197–1203, 2015.

[8] H. Azami, A. Fernández, and J. Escudero, "Refined multiscale fuzzy entropy based on standard deviation for biomedical signal analysis," *Medical & Biological Engineering & Computing,* vol. 55, pp. 2037–2052, 2017.

[9] T. J. Licznerski, J. Jaroński, and D. Kosz, "Ultrasonic system for accurate distance measurement in the air," *Ultrasonics,* vol. 51, pp. 960–965, 2011.

[10] A. L. Goldberger, C.-K. Peng, and L. A. Lipsitz, "What is physiologic complexity and how does it change with aging and disease?," *Neurobiology of aging,* vol. 23, pp. 23–26, 2002.

[11] A. I. Dastiridou, H. S. Ginis, D. De Brouwere, M. K. Tsilimbaris, and I. G. Pallikaris, "Ocular rigidity, ocular pulse amplitude, and pulsatile ocular blood flow: the effect of intraocular pressure," *Investigative Ophthalmology & Visual Science,* vol. 50, pp. 5718–5722, 2009.

 ISBN: 978-1-61208-716-0

Chaos-Based Communication Systems Based on the Sprott D Attractor

Carlos E. C. Souza
Federal University of Pernambuco
Recife, Brazil 50740-550
Email: carlos.ecsouza@ufpe.br

Daniel P. B. Chaves
Federal University of Pernambuco
Recife, Brazil 50740-550
Email: daniel.chaves@ufpe.br

Cecilio Pimentel
Federal University of Pernambuco
Recife, Brazil 50740-550
Email: cecilio@ufpe.br

***Abstract*—We propose a chaos-based communication system employing the symbolic dynamics of the Sprott D chaotic attractor. The chaotic dynamics is modeled by a graph representing the evolution of the chaotic trajectories. Finite-state encoders are employed to map unrestricted binary sequences in the restricted sequences generated by the discretization of the chaotic flow. Finally, the performance of the proposed system is analyzed.**

Keywords–Chaos communication; Symbolic dynamics; Topology of three-dimensional chaotic attractors.

I. INTRODUCTION

Chaos-based communication systems are suitable alternatives in modern communications due to some inherent characteristics of the chaotic behavior such as decorrelation, non-periodic behavior, broadband spectrum, energetic efficiency [1]. The use of chaotic dynamical systems has been considered in several different scenarios [2]. In particular, three-dimensional chaotic attractors have been successfully employed in the design of digital communication systems [3].

The chaotic trajectories within three-dimensional chaotic attractors can be discretized by a labeled partition of a Poincaré section. This defines a map between the continuous flow and the symbolic representation of the chaotic system. The symbolic sequences generated by a chaotic attractor are restricted due to the dynamical constraints imposed by the chaotic flow, and these restrictions can be exploited in the design of chaos-based communication systems. The dynamical mechanism of a chaotic attractor is represented by a set $\mathcal{F}$ that represents *words* (or sequences) that never occur in the symbolic representation of the attractor dynamics. The forbidden set $\mathcal{F}$ is used to construct a graph that represents the discrete dynamics of the system and is employed in the design of Finite-State Encoders (FSE), used to map unrestricted into restricted sequences. Moreover, the FSE are used in the decoder to estimate the information sequence.

II. MAIN RESULTS

To illustrate the proposed methodology, we detail the design of a communication system based on the Sprott D system, defined by

$$\begin{cases} \dot{x}(t) &= -y(t) \\ \dot{y}(t) &= x(t) + z(t) \\ \dot{z}(t) &= x(t)z(t) + ay^2(t) \end{cases} \quad (1)$$

where $a = 2.3$ is the control parameter [4]. The Poincaré section for the Sprott D system is a plane perpendicular to the xy plane with $y = 0$. Figure 1 shows the Sprott D attractor in projection on the xy plane and its Poincaré section with a labeled binary partition over the alphabet $\mathcal{A}_2 = \{\mathrm{A}, \mathrm{B}\}$. The threshold of the partition is defined by the minimum of the Poincaré return map. When a chaotic trajectory crosses the Poincaré section, we assign the label of the region where the crossing occurs in the symbolic evolution of the system. The continuous flow is mapped into symbolic sequences $S_N = s_0 s_1 s_2 \dots s_{N-1}$, where $s_i \in \mathcal{A}_2$ is the label of the visited region in the Poincaré section in the ith crossing. Therefore, S_N records the order of visitation in the regions of the Poincaré section in N consecutive crossings.

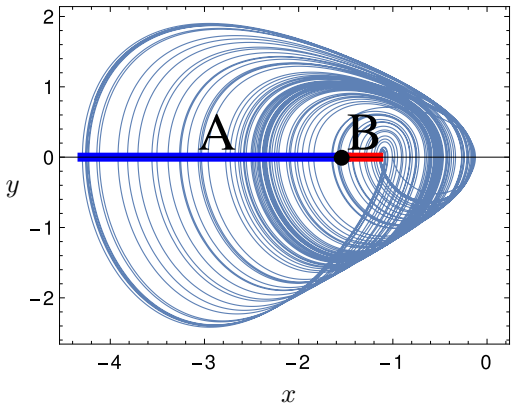

Figure 1. Sprott D attractor and its Poincaré section with a labeled binary partition in projection on the xy plane.

The first step in the proposed methodology is to find the set $\mathcal{F}$ for the Sprott D system. We search in S_N for words with length n that never occur as sub-sequences of S_N. In this work, we consider $n \leq 10$ and in this case the set $\mathcal{F}$ is

$$\mathcal{F} = \{\mathrm{BB}, \mathrm{BAAB}, \mathrm{BAAAAB}\}. \quad (2)$$

Applying the procedures described in [5], we construct a graph that represents the dynamics of a restricted system according to the specification of the set $\mathcal{F}$. Then, we construct the FSE for the Sprott D system employing the method described in [6], which is illustrated in Figure 2. The FSE maps one information bit into two symbols $s_1 s_2$. The next step to design the communication system is to associate the chaotic signals with the state transitions of the FSE.

The states transitions $s_1 s_2$ of the FSE in Figure 2 correspond to two successive crossings in the Poincaré section

 ISBN: 978-1-61208-716-0

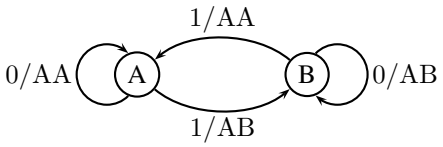

Figure 2. Finite-state encoder for the Sprott D attractor.

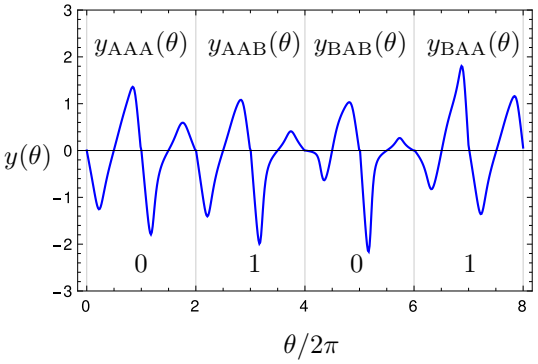

Figure 3. Chaotic signals generated by the Sprott D attractor versus the angle θ to encode the information sequence 0101.

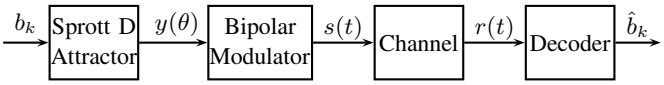

Figure 4. Block diagram of the communication system based on the Sprott D chaotic attractor.

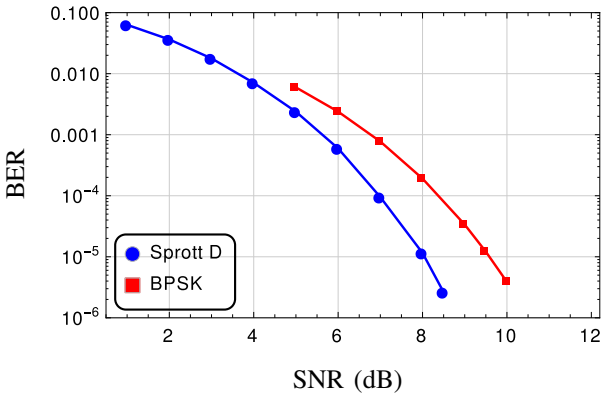

Figure 5. BER for the Sprott D communication system over an AWGN channel. The BER for an uncoded BPSK is also shown.

illustrated in Figure 1. The chaotic signals associated with the transitions are defined as the segments of the chaotic trajectories generated by the variables of the system connecting the corresponding regions of the Poincaré section. Any variable can be used and in this work we employ the variable y. We define the angle θ as the angle between the vector obtained by connecting the center of the attractor to a point of a chaotic trajectory in projection on the xy plane and the Poincaré section [3]. Parameterizing the chaotic signals as $y(\theta)$, the crossings are periodic with angular period 2π. For example, the chaotic signal associated with the transition from state A to state B in the FSE is the segment of trajectory $y(\theta)$ connecting the regions A to A and subsequently connecting the regions A to B, denoted $y_{\mathrm{AAB}}(\theta)$, with angular period 4π. Fig. 3 illustrates a chaotic trajectory versus the angle θ that encodes the information sequence 0101 starting at state A.

The proposed communication system is represented in Figure 4. The state transitions are induced by chaos control [7] according to the information bits $b_0 b_1 b_2 \ldots$. The chaotic waveforms $y(\theta)$ are inverted (multiplied by -1) in the bipolar modulator when the information bit is 0 [3]. At each signaling interval T, the signal $y(\theta)$ is mapped into the transmitted signal $s(t)$ with duration T. The signal $s(t)$ is transmitted over an AWGN channel and the received signal is $r(t) = s(t) + n(t)$, where the noise $n(t)$ has uniform power spectral density $N_0/2$. The decoder uses the Viterbi algorithm to estimate the information sequence.

We performed numerical simulations to evaluate the Bit Error Rate (BER) of the proposed communication system. The Signal to Noise Ratio (SNR) is defined as $\mathrm{SNR} = \bar{E}_s/N_0$, where $\bar{E}_s$ is the average energy of the transition curves. Figure 5 shows the performance of the proposed system and the uncoded BPSK system for reference purposes. The proposed system outperforms the BPSK system by approximately 1.8 dB for a BER $= 10^{-5}$.

III. Conclusions

We proposed a methodology to design chaos-based communication systems using a discretization of the chaotic flow generated by the Sprott D chaotic attractor. The discrete chaotic dynamics was used to design an FSE to transmit information when the dynamics is restricted, resulting in better robustness against the channel noise. Therefore, the proposed system is an interesting alternative in chaos-based communications, yielding enhanced performance at the cost of a relatively simple encoding scheme.

Acknowledgment

The authors would like to thank CNPq and FACEPE for the financial support.

References

[1] S. Strogatz, Nonlinear Dynamics and Chaos with Applications to Physics, Biology, Chemistry, and Engineering, ser. Studies in Nonlinearity Series. Westview Press, 2001.

[2] G. Kaddoum, "Wireless chaos-based communication systems: A comprehensive survey," IEEE Access, vol. 4, May 2016, pp. 2621–2648.

[3] C. E. C. Souza, D. P. B. Chaves, and C. Pimentel, "Digital communication systems based on three-dimensional chaotic attractors," IEEE Access, vol. 7, Jan. 2019, pp. 10 523–10 532.

[4] J. C. Sprott, "Some simple chaotic flows," Phys. Rev. E, vol. 50, Aug. 1994, pp. R647–R650.

[5] M. Crochemore, F. Mignosi, and A. Restivo, "Automata and forbidden words," Information Processing Letters, vol. 67, no. 3, Aug. 1998, pp. 111 – 117.

[6] D. Lind and B. Marcus, An Introduction to Symbolic Dynamics and Coding. Cambridge University Press, 1995.

[7] S. Hayes, C. Grebogi, and E. Ott, "Communicating with chaos," Phys. Rev. Lett., vol. 70, May 1993, pp. 3031–3034.

 ISBN: 978-1-61208-716-0

Integrated Streak Camera With on Chip Averaging for Signal to Noise Ratio Improvement

Wilfried Uhring, Jean-Baptiste Schell, Luc Hébrard
ICube laboratory
University of Strasbourg and CNRS
Strasbourg, France
Emails: wilfried.uhring@unistra.fr, jbschell@unistra.fr, luc.hebrard@unistra.fr

***Abstract*— A technique to reject the noise of the Trans-Impedance Amplifiers (TIAs) of an integrated streak camera in the case of a repeatable input signal is proposed. The modifications of the sensor architecture are marginal with only one additional power supply connected to the column buffer of the sampling cell. The noise rejection can be adjusted independently of the effective bandwidth of the system. The simulation results show that the TIA noise can be reduced from 3.5mV to 0.31mV, which is the limit of the used sampling cells due to the thermal noise on capacitor (kTC). The resulting signal-to-noise ratio is more than 10 times better with an acquisition time of only 20 µs thanks to the on-chip analog averaging feature.**

Keywords—High-speed imaging; CMOS; N-path filter; noise.

I. INTRODUCTION

High-speed imaging using Complementary Metal-Oxide Semiconductor (CMOS) or Charged Coupled Device (CCD) sensors has seen a sharp increase with the apparition of commercial products able to achieve a frame rate of 20 Mega frames per second (Mfps) [1][2]. Thanks to the concept of burst imaging introduced in 1993 by Pr. Etoh Goji [3], this sensor can generate a pixel rate in the range of 1 Tera pixels per second by storing the images within the sensor during the acquisition. Recently, the 3D microelectronic technology allowed to push the state of the art of the 2D high speed burst video sensors with some new functionalities such as an on-the-fly digitalization [4] and an enhanced frame rate of 100 Mfps [5][6]. Nevertheless, in order to reach a higher frame rate, the 2D acquisition paradigm has to be abandoned to the benefit of the streak imaging paradigm. The streak imaging approach is the sampling of just a single spatial line of the scene per unit of time. As a consequence, a spatial dimension is lost and the video is no longer a pile of spatial images *I(x,y)* where *x* and *y* are the spatial dimensions, but a pile of lines which can be represented as a spatiotemporal image *I(x,t)*, where *t* is the time. Indeed, the fastest frame rate achieved with a CMOS sensor has been realized thanks to a streak imaging sensor, such as the one shown in Fig. 1 [7]. By releasing the silicon area constraints of the 2D imagers, one single pixel (the grayed line of Fig. 1) can set in a front end with a large sampling and storage unit beside the photodiode, with almost no spatial limit except the sensor width. Thus, a wideband front end and sampling unit can be embedded, and a line rate of 8 Giga lines per second with a temporal resolution better than 500 ps has been demonstrated [7].

The drawback of such a high bandwidth performance is the noise of the system that increases with the cut-off frequency. Although a high bandwidth is required for single shot measurements, there are some other approaches to measure a repetitive event. Time correlated single photon counting is obviously a very good technique to measure the temporal evolution of a light pulse with a very high signal-to-noise ratio, but the acquisition time can be quite long and these sensors are intrinsically unable to achieve a single shot measurement [8]. In this paper, we propose a technique to reduce the noise of a single shot integrated streak camera with very few modifications of the architecture.

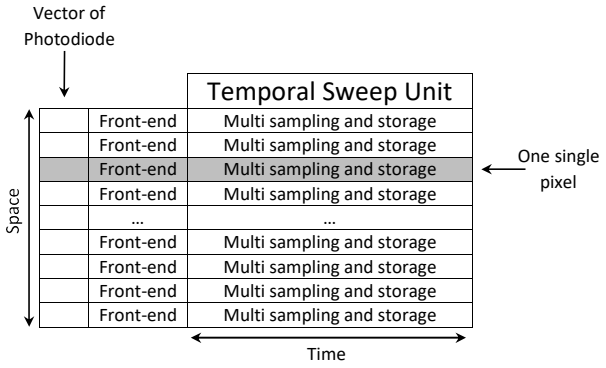

Fig. 1. A streak camera sensor architecture

The overall system architecture and theory of operation are detailed in Section II. The approach functionality and efficiency are then assessed with simulation results depicted in Section III. Section IV concludes on the work.

II. SYSTEM ARCHITECTURE AND OPERATION

A. System architecture

A pixel of the integrated streak camera is composed of a photodiode, a wideband Trans-Impedance amplifier (TIA) and a sampling and storage line, which is a bank of capacitors switched with NMOS (N-type Metal Oxide Semiconductor) transistors, as described in Fig. 2. Each NMOS switch is driven by the S_i signal generated by the temporal sweep unit with $i \in [1;N]$ and N the number of images of the line stored in the sensor. The S_i signals are set on and off sequentially in order to sample the signal applied on the common line by the TIA *A*, which gives an image of the light intensity on the photodiode. After the sampling of the signal, the voltage C_i on each holding capacitor C_{Hi} is then readout by a source follower. Both the photodiode and the readout unit are not presented in Fig. 2 for clarity reason, but they are detailed in [7].

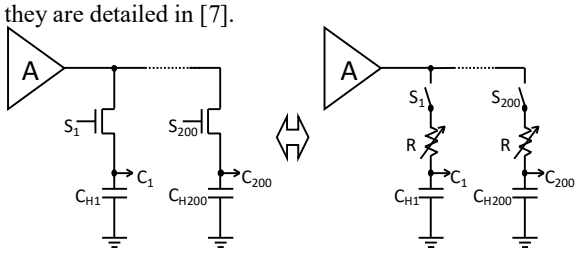

Fig. 2. Simplified architecture of a single pixel and its equivalent model

 ISBN: 978-1-61208-716-0

The bandwidth of the sampling cell can be well above the bandwidth of the TIA which is about 1 GHz. Thus, each time the NMOS is switched on, we can assume that the voltage on the node C_i follows instantaneously the TIA output. This is the operation in the single shot mode. Nevertheless, the noise at the output of the TIA is also sampled and affects the signal-to-noise ratio.

B. Theory of operation of the on chip averaging

It is well known that averaging a signal reduces the temporal noise and, if the signal can be repeated, multiple acquisitions of a single shot can be readout and averaged on a computer. However, the repetition rate of the event is then limited by the readout time, i.e., about a few Hz. This bottleneck can be circumvented by a high repetition rate on chip averaging. By reducing the voltage of the gate of the NMOS transistor, the "on resistance" R of the transistor can be increased and the behavior of the sampling cell is equivalent to an ideal switch followed by a resistor and the holding capacitance, as depicted in Fig. 2.

If the RC_{Hi} time constant of the sampling cell is higher than the switch on duration T of the S_i signal, the voltage on the node C_i does not have time to follow the output of the TIA. At the same time, the TIA noise is also filtered by the RC_{Hi} filter. When the camera is synchronously triggered by the laser source or vice versa, each time the signal is repeated, the node voltage C_i tends to reach the signal during the switch on duration T. In the meantime, the temporal noise is still filtered and is randomly positive or negative. As a consequence, a high noise rejection can be achieved by reducing the sampling cell bandwidth. The system behaves like an N-Path filter and the transfer function can be written as [9]:

$$\frac{V_{out}(s)}{V_{in}(s)} = \sum_{K=-\infty}^{K=+\infty} \mathrm{sinc}^2\left(\frac{K\pi}{N}\right) \cdot G\left(s - jK\omega_0\right),\ K \in \mathbb{Z} \qquad (1)$$

where $V_{in}(s)$ is the signal at the output of the TIA, $V_{out}(s)$ is the signal reconstructed with the readout of the C_i node voltage and $G(s)=1/(1+RC_H \cdot s)$ is the transfer function of the RC_H filter. Equation (1) is equivalent to a comb Dirac modulated by a sinc due to the gate function of the S_i signal and convolved to the elementary filter $G(s)$.

The complete transfer function from (1) is illustrated in Fig. 3 for sampling cell bandwidths of 100 kHz and 10 kHz for N=200 cells and a sampling rate of 1 GHz, i.e., a period T of 1 ns. As the signal is periodically repeated, it can be decomposed in a Fourier series with fundamental frequency at 1 GHz/200, i.e., 5 MHz and its harmonics. Thus, the complete transfer function for the measured signal is almost equal to 1 for its fundamental and is slightly modulated by the sinc function for its harmonics. Indeed, each sample trends to be the mean voltage of the TIA signal along the gate duration T of the sampling signal which is the reason of the low pass behavior of the sinc function.

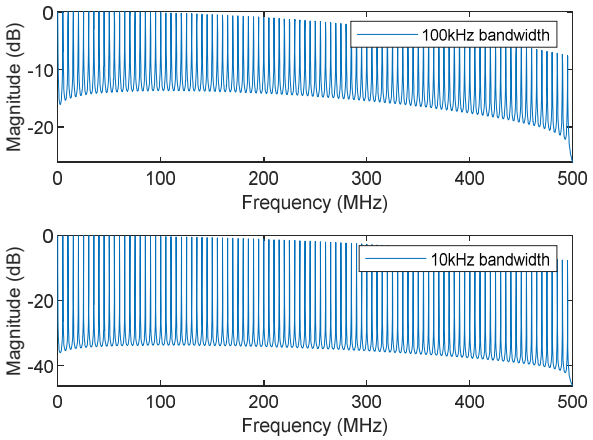

Fig. 3. Complete transfer function of the on-chip averaging system for sampling cell bandwidths of 100 kHz (top) and 10 kHz (bottom)

C. Noise rejection

Fig. 4 is a close-up view of the complete transfer function at the frequency of 5 MHz for sampling cell bandwidths of 100 kHz and 10 kHz. Its magnitude at 5 MHz is actually 0 dB in order for the signal fundamental to fully pass through the filter. The other spectral peaks of the complete transfer function look similar, except for their magnitude modulated by the sinc function.

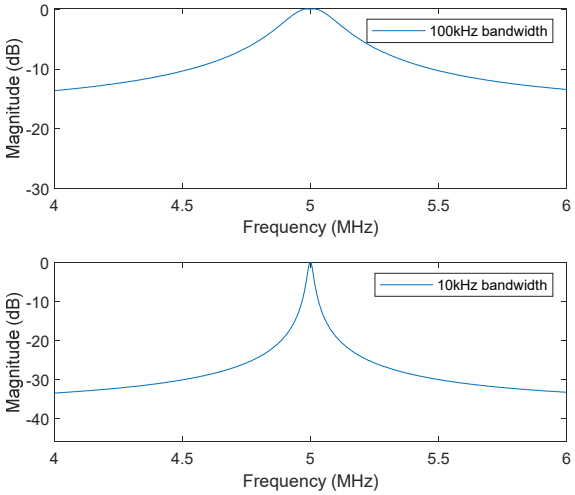

Fig. 4. Close-up view of the fundamental at 5 MHz of the complete transfer function of the one chip averaging system for sampling cell bandwidths of 100 kHz (top) and 10 kHz (bottom)

If we assume that the temporal noise at the output of the TIA is a white noise, this noise is rejected between two spectral peaks thanks to the low pass filter G(s). The narrower are the peaks, the higher is the rejection. Nevertheless, the overlap of the response of two peaks limits the maximal rejection. Indeed, the transfer functions given in Fig. 3 and Fig. 4 indicate a rejection of about 15 and 30 dB for a filter bandwidth of respectively 100 kHz and 10 kHz. With an ideal brickwall filter, the noise rejection should be increased by √n where n is the ratio of the filter bandwidth, i.e., the rejection should be √10 higher with a filter bandwidth of 10 kHz with respect of the one at 100 kHz.

TABLE I. NOISE REJECTION IMPROVEMENT FOR SEVERAL FILTER BANDWIDTH

100 kHz ➔ 10 kHz	10 kHz➔1 kHz	1 kHz➔100 Hz
4.25	3.36	3.1

TABLE I. gives the computed white noise rejection for the transfer function (1) with a first order filter G(s) = 1/(1+RCs) and a bandwidth of 100 kHz, 10 kHz and

1 kHz respectively. The noise rejection is close to √10=3.16 for a decade of bandwidth.

D. Practical limitations

Lowering the filter bandwidth rejects the noise but, as a drawback, a higher number of signal repetition is required. Indeed, a sampling cell is switched on only during the sampling duration T, i.e., the duty cycle is nominally 1/N. As a consequence, the apparent constant time τ_a of the cell is about N times the constant time of the RC_H filter, i.e., $\tau_a=N{\cdot}R{\cdot}C_H$. Moreover, the sampling cell suffers from a leakage current that makes it unable to operate with a too long time constant. In a conventional CMOS technology, the leakage time constant is in the range of 100 ms. Thus, the minimal usable filter bandwidth is in the range of 1 kHz.

The noise rejection is also limited by kTC noise which is the thermal noise of the resistance R integrated on the capacitance and is given by (2).

$$V_{ktc} = \sqrt{\frac{k \cdot T}{C_H}} \qquad (2)$$

III. SIMULATION RESULTS

A. Bandwidth tunning

The bandwidth of cell filter $G(s)$ can be adjusted by adjusting the gate to source voltage V_{GS} of the NMOS transistor in order to obtain an on resistance R that matches $f=1/(2{\cdot}\pi{\cdot}R{\cdot}C_H)$ where f is the required bandwidth and C_H is the value of the holding capacitance. The on resistance follows a strongly nonlinear response according to V_{GS} that allows generating a large dynamic of resistance value R and thus, a large dynamic of bandwidth. The Gate voltage is generated thanks to buffers with an adjustable power supply. Though the source of the transistor is following the TIA output and consequently, the V_{GS} is signal dependent. Fig. 5 gives the small signal bandwidth of the sampling cell filter versus the buffer power supply voltage for different TIA output voltages for the designed sampling cell. More than 3 decades can easily be covered thanks to this technique which does not require any additional transistors within the sampling cell. The only modification relies on the tuning of the supply voltage of the buffer that drives the S_i voltage. These buffers are common for a whole column of the sensor and each column buffer is powered by the same power supply. As a consequence, the only modification of the system architecture is to add a specific power supply for these buffers.

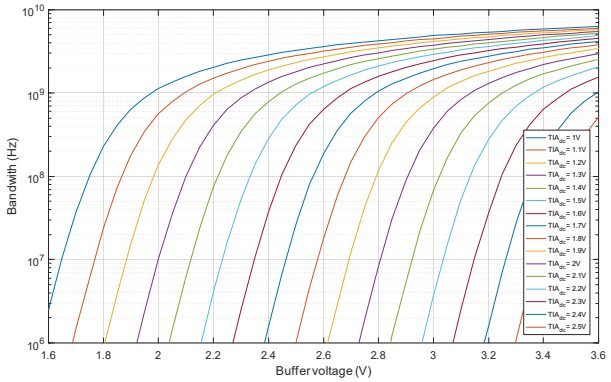

Fig. 5. Small signal bandwidth of the sampling cell filter versus the buffer power supply voltage for different TIA output voltage

B. Sinusoidal signal acquisition

The noise rejection has been evaluated with a SPICE simulation for two sinusoidal signals of 5 and 50 MHz at the input of the TIA. The total added rms noise at the output of the TIA integrate from 10 Hz to 10 GHz is 3.5mV rms. The results are shown in Fig. 6 for several filter bandwidth of 1 GHz down to 1 MHz. The displayed curves are the concatenation of the voltage C_i from the sampling cells i=1 to 200 after 100 accumulations of the same signal. The simulated time is then 200 cells×1 ns×550 accumulations = 110 µs.

With a filter bandwidth of 1 GHz (pink curve of Fig. 6), the sampled signal with 100 accumulations is almost the image of the last one. Indeed, during the 1 ns long aperture of the switches, the voltage at the nodes C_i has enough time to reach the TIA output. In this case, the results are very close to a classical sampling and we can see that the noise added by the TIA is present on the sampled signal.

With a filter bandwidth of respectively 100 MHz (green), 10 MHz (blue) and 1 MHz (red), the noise is less and less present on the sinusoids. We can also see that the shape and amplitude of the sinusoids are not affected by the filter bandwidth reduction and that the 50 MHz signal can be measured even with a filter bandwidth of only 1 MHz. This simulation demonstrates that the white noise can be efficiently rejected without altering the periodic signal.

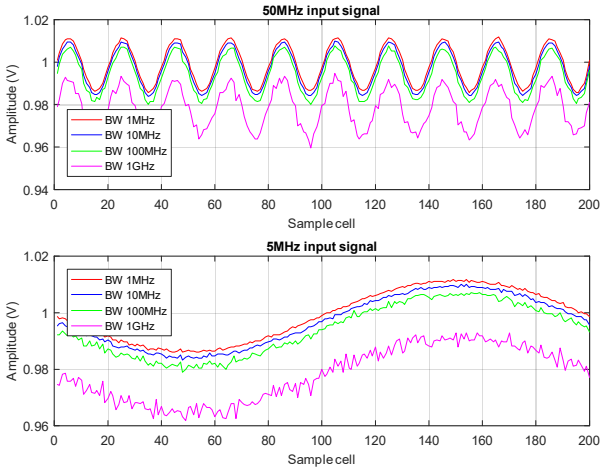

Fig. 6. Simulation of the noise rejection on a sinusoidal input signal of 50 MHz (top) and 5 MHz (bottom) for a filter bandwidth of 1 GHz down to 1 MHz. A cell is equivalent to 1 ns

C. Noise rejection assessment

In order to characterize the noise rejection, the same simulation as the previous one was made without any signal at the input. The observed signal should be the operating point of the TIA output, i.e., a static voltage of about 1 V. The result is shown in Fig. 7.

Once again, we can clearly see that the noise is rejected and that the reconstructed signal becomes less noisy. The equivalent power spectral densities of the measured noises are plotted in Fig. 7. We can see that the rejection is applied on the whole spectrum of the white noise and especially also for the low frequencies. The statistical distributions of theses samples are also depicted in Fig. 8. The assessed noise with a bandwidth of 1 GHz is 1.75 mV rms and is reduced down to 0.769 mV rms for a bandwidth of 100 MHz, i.e., a

 ISBN: 978-1-61208-716-0

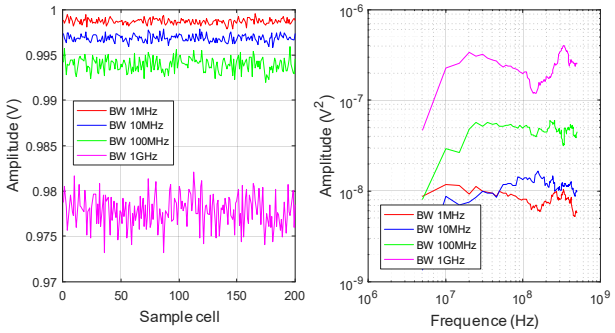

Fig. 7. Simulation of the TIA white noise rejection for different filter bandwidth. Acquired Samples (left) and equivalent power spectral density (right)

rejection of 2.3. For a bandwidth of 10 Mhz, the noise is reduced again down to 0.383 mV rms, i.e., the rejection from 100 to 10 MHz is about 2. Finally, the rejection over the decade from 10 to 1 MHz is only 1.2 with a noise at 1 MHz of 0.31 mVrms. This last simulated rejection ratio is less than expected because it is limited by the kTC noise of the sampling cell which integrates a holding capacitance of 40 fF, i.e., $V_{kTC} = 0.32$ mV. The proposed noise rejection technique makes the TIA noise negligible with respect to the kTC noise level.

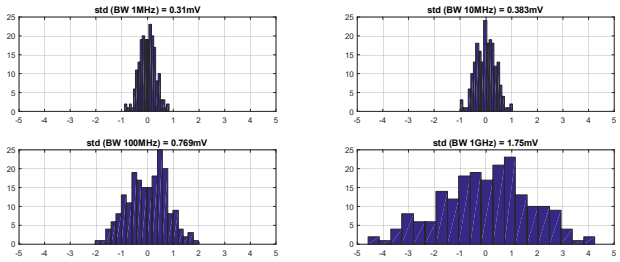

Fig. 8. Noise distribution (mV) of the simulated sample for noise rejection assessment for different filter cell bandwidths

IV. Conclusion

A technique to reject the noise of the TIAs of an integrated streak camera in the case of a repeatable input signal is proposed. The modifications of the sensor architecture are marginal with only one additional power supply connected to the column buffer of the sampling cell. The noise rejection can be adjusted independently of the effective bandwidth of the system. The simulation results show that the TIA noise can be reduced from 3.5mV to 0.31mV which is the limit of the used sampling cells due to the kTC noise. The resulting signal-to-noise ratio is more than 10 times better with an acquisition time of only 20 µs thanks to the on-chip analog averaging features.

References

[1] High-speed video camera, Shimadzu model Hyper Vision HPV-X2, https://www.shimadzu.com/an/test/hpv/index.html, retrieved: April, 2019

[2] Ultra_High speed Video, Specialised Imaging, Model Kirana, https://specialised-imaging.com/products/kirana-high-speed-video-camera , retrieved: April, 2019

[3] T. Goji Etoh, K. Takehara, and K. Takehara, "Trahigh-speed multiframing camera with an automatic trigger," Proc. SPIE 1757, Ultrahigh- and High-Speed Photography, Videography, and Photonics, (19 January 1993), doi: 10.1117/12.139154.

[4] L. Millet et al. "A 5 Million Frames Per Second 3D Stacked Image Sensor With In-Pixel Digital Storage," ESSCIRC, 44th European Solid-State Circuits Conference, Dresden, Germany, 2018, pp. 62-65, doi:10.1109/ESSCIRC.2018.8494287.

[5] R. Kuroda, M. Suzuki, and S. Sugawa, "Over 100 million frames per second high speed global shutter CMOS image sensor," Proc. SPIE 11051, 110510B (28 January 2019), doi: 10.1117/12.2524492.

[6] W. Uhring et al., "A Scalable Architecture for Multi Millions Frames per Second CMOS Sensor With Digital Storage," IEEE NEWCAS 2018, Montréal, Canada, 2018, pp. 252-255, doi:10.1109/NEWCAS.2018.8585644.

[7] M. Zlatanski, W. Uhring, and J-P. Le Normand, "Sub-500 ps Temporal Resolution Streak-mode Optical Sensor," IEEE Sensors Journal, Institute of Electrical and Electronics Engineers (IEEE), pp. 6570-6583, 15, n° 11, 2015, doi:10.1109/JSEN.2015.2462021.

[8] W. Becker, "Advanced Time-Correlated Single Photon Counting Techniques," Part of the Springer Series in Chemical Physics book series, Vol. 81, 2005, Eds. Springer, ISBN-13 978-3-540-26047-9

[9] A.L. Jones, "Theory and Performanc of N-Path Filters," Technical report 3602-1, Standford electronics laboratories, 1971.

 ISBN: 978-1-61208-716-0