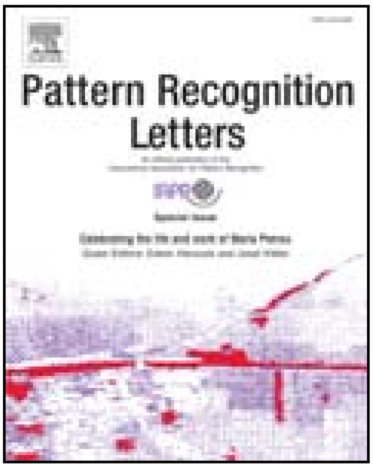
Accepted Manuscript



Simultaneous Robustness against Random Initialization and Optimal Order Selection in Bag-of-Words Modeling

Ibrahim F. Jasim Ghalyan, Sonia M. Chacko, Vikram Kapila

PII: S0167-8655(18)30580-4
DOI: https://doi.org/10.1016/j.patrec.2018.09.010
Reference: PATREC 7309

To appear in: *Pattern Recognition Letters*

Received date: 3 May 2018
Revised date: 9 August 2018
Accepted date: 7 September 2018

Please cite this article as: Ibrahim F. Jasim Ghalyan, Sonia M. Chacko, Vikram Kapila, Simultaneous Robustness against Random Initialization and Optimal Order Selection in Bag-of-Words Modeling, *Pattern Recognition Letters* (2018), doi: https://doi.org/10.1016/j.patrec.2018.09.010

Pattern Recognition Letters

Authorship Confirmation

Please save a copy of this file, complete and upload as the "Confirmation of Authorship" file.

As corresponding author I, Vikram Kapila, hereby confirm on behalf of all authors that:

1. This manuscript, or a large part of it, has not been published, was not, and is not being submitted to any other journal.
2. If presented at or submitted to or published at a conference(s), the conference(s) is (are) identified and substantial justification for re-publication is presented below. A copy of conference paper(s) is(are) uploaded with the manuscript.
3. If the manuscript appears as a preprint anywhere on the web, e.g. arXiv, etc., it is identified below. The preprint should include a statement that the paper is under consideration at Pattern Recognition Letters.
4. All text and graphics, except for those marked with sources, are original works of the authors, and all necessary permissions for publication were secured prior to submission of the manuscript.
5. All authors each made a significant contribution to the research reported and have read and approved the submitted manuscript.

Signature____ Date_______________________

List any pre-prints:

Relevant Conference publication(s) (submitted, accepted, or published):

Justification for re-publication:

Graphical Abstract (Optional)

To create your abstract, please type over the instructions in the template box below. Fonts or abstract dimensions should not be changed or altered.

Simultaneous Robustness against Random Initialization and Optimal Order Selection in Bag-of-Words Modeling

Ibrahim F. Jasim Ghalyan
Sonia M. Chacko
Vikram Kapila

This article proposes an Enhanced Stochastically Robust and Optimized Bag-of-Words (ESRO-BoW) modeling technique that simultaneously accounts for the problems of robustness against random initialization and optimal model-order selection in BoW modeling. To address the aforementioned problems, the modeling performance of multiple executions of the BoW technique is considered as a discrete random variable and the ESRO-BoW is developed such that a convergence in mean is guaranteed for the resulting sequence of random variables. The BoW model order is tuned such that the expected value of the limit of the random variable of the classification performance is maximized. Hence, the ESRO-BoW realizes both robustness against random initializations and selects the optimal BoW model order. In order to evaluate its efficiency, the ESRO-BoW is applied to the classification of Caltech 101 image set and excellent performance is obtained. Comparison with the state-of-the-art approaches, employed for classifying Caltech 101 image set, demonstrates the superiority of the suggested ESRO-BoW modeling technique.

Research Highlights (Required)

To create your highlights, please type the highlights against each `\item` command.

It should be short collection of bullet points that convey the core findings of the article. It should include 3 to 5 bullet points (maximum 125 characters, including spaces, per bullet point.)

- An enhancement to Bag-of-Words (BoW) is proposed to produce robustness against random initialization and optimal model-order.
- Robustness against random initialization is realized by treating performance as a random variable and estimating its limit.
- The optimal model order is estimated such that the expected value of the performance is maximized.
- The suggested BoW enhancement is applied to the Caltech 101 image set classification and excellent performance is obtained.
- Comparison is conducted with the available state-of-the-art techniques and the superiority of the ESRO-BOW is demonstrated.

Pattern Recognition Letters
journal homepage: www.elsevier.com

Simultaneous Robustness against Random Initialization and Optimal Order Selection in Bag-of-Words Modeling

Ibrahim F. Jasim Ghalyan[a], Sonia M. Chacko[a], and Vikram Kapila[a,**]

[a] *Mechatronics, Controls, and Robotics Laboratory, Tandon School of Engineering, New York University, Six Metrotech Center, Brooklyn NY 11201, USA*

ABSTRACT

This article proposes an Enhanced Stochastically Robust and Optimized Bag-of-Words (ESRO-BoW) modeling technique that simultaneously accounts for the problems of robustness against random initialization and optimal model-order selection in BoW modeling. To address the aforementioned problems, the modeling performance of multiple executions of the BoW technique is considered as a discrete random variable and the ESRO-BoW is developed such that a convergence in mean is guaranteed for the resulting sequence of random variables. The BoW model order is tuned such that the expected value of the limit of the random variable of the classification performance is maximized. Hence, the ESRO-BoW realizes both robustness against random initializations and selects the optimal BoW model order. In order to evaluate its efficiency, the ESRO-BoW is applied to the classification of Caltech 101 image set and excellent performance is obtained. Comparison with the state-of-the-art approaches, employed for classifying Caltech 101 image set, demonstrates the superiority of the suggested ESRO-BoW modeling technique.

1. Introduction

Deep learning has recently emerged as one of the most promising techniques to perform data-driven modeling, e.g., image classification, with impressive performance (see Lecun et al. (2015) and references therein). However, deep learning has been shown to be susceptible to multiple issues, e.g., weak interpretability (Nguyen et al., 2015), requirement of large data set (Camilleri and Prescott, 2017), sensitivity to imperfection in the training (Papernot et al., 2016), among others. To avoid such limitations, alternative scalable methods are frequently utilized for data-driven modeling and the Bag-of-Words (BoW) is considered one of the efficient interpretable modeling methods. Due to its excellent performance, the BoW modeling technique has attracted the interest of many researchers and practitioners who have successfully employed it in fields such as text recognition, computer vision, automation, heath-care, among others. It consists of three key steps, feature extraction, quantization, and classification, which are integrated to produce an efficient technique applicable for various image classification problems (Csurka et al., 2004).

In Lazebnik et al. (2006), spatial pyramid matching (SPM) was introduced in the BoW modeling where images are partitioned into successively increasing fine sub-regions and histograms of local features, inside the sub-regions, are computed. The visual words, obtained in Lazebnik et al. (2006), are assumed to be uniformly distributed. To relax the need for nonlinear classifiers with SPM, locality-constrained linear coding (LLC) was introduced in which each descriptor is projected on its local coordinate system leading to possible employment of linear classifiers, simplifying computations encountered in the case of nonlinear classifiers (Wang et al., 2010). The excellent performance reported using the BoW modeling method spurred many researchers to employ this modeling technique in classifying human motion (Burghouts and Schutte, 2013; Hernández-Vela et al., 2014; Iosifidis et al., 2014). Based on the local features of the song level, BoW was efficiently utilized to classify music sounds and promising results were obtained (Fu et al., 2011). In dos Santos et al. (2015); Karakasis et al. (2015), the BoW modeling technique was employed to produce an efficient image retrieval process. By considering a set of words, in which the order of the words is taken into account, the BoW was en-

[**] Corresponding author: Tel.: +1-646-997-3161; fax: +1-646-997-3532; *e-mail:* vkapila@nyu.edu (and Vikram Kapila)

hanced to handle complex situations like accommodating possible scene occlusion (Bolovinou et al., 2013).

In Wu and Wong (2012), the BoW was efficiently employed for segmenting and tracking multiple objects moving in a scene by extracting, from the optical flow, the field pattern representing the collective movement of objects. A promising classification technique between machine-printed and handwritten texts was obtained by Zagoris et al. (2014) using the BoW modeling technique that was coupled with an already trained and optimal available codebook. Multisupport region order-based gradient histogram (MROGH) technique has been integrated in the BoW modeling approach by Shen et al. (2014) for classifying human epithelial type 2 (HEp-2) cells. Specifically, the MROGH enhanced the process of feature extraction when performing the BoW and the overall results were promising. In Mu et al. (2015), LLC is combined with spatial distribution pooling (SDP) where the distribution of visual words is characterized by Gauss mixture model (GMM), rather than the uniform distribution considered in Lazebnik et al. (2006), and improved BoW modeling is reported. Additional developments have been reported in the theory and application of the BoW modeling process such as neural network-based BoW (Passalis and Tefas, 2017), using the group of local words to develop Bag-of-Expressions (BoE) (Nazir et al., 2018), among others.

In spite of the advances in BoW modeling and its application, the problem of random initialization in BoW modeling has not been addressed yet. In fact, this problem can lead to solutions potentially getting trapped in local minima during the implementation of the BoW steps, thus causing performance degradation in the BoW modeling process. Moreover, poor order selection of the BoW modeling can result in either models that are not sufficiently representative of the phenomenon, if a relatively small order is used, or overfitting, in the case of a large order. Although a trial and error process is frequently followed in estimating the BoW model order, it does not provide an optimized model order for a given set of data, thus failing to reveal the highest obtainable performance of the BoW modeling process. Moreover, despite the work reported in Yang et al. (2007) concerning the relationship between the classification performance and the BoW model order (i.e., the number of visual words of the model), automating the estimation process for model order to produce optimal performance also remains to be investigated.

In this paper, we propose an Enhanced Stochastically Robust and Optimized Bag-of-Words (ESRO-BoW) modeling technique to address both the problems of random initialization and model order selection. Thus, the main contributions of the ESRO-BoW modeling technique of this paper are to simultaneously:

1. Develop models that are robust against random initialization of parameters.
2. Estimate the optimal model order that yields the optimal modeling performance.

The performance of the BoW modeling is considered as a random variable with robustification and optimization of this performance as the main objective of this paper. Section 2 summarizes the BoW modeling technique and several important preliminary concepts are reviewed in section 3. Development of the ESRO-BoW modeling technique is detailed in section 4 while the experimental validation is explained in section 5. Finally, section 6 provides some concluding remarks.

2. Bag-of-Words (BoW) Modeling Scheme

The BoW modeling process is composed of three main steps: feature extraction, quantization, and classification. Below is a brief description of these steps.

- **Feature Extraction:** Feature extraction is the first step of the BoW and it consists of two sub-steps. In the first substep the interest points of the image are detected using any available technique such as scale-invariant feature transform (SIFT), speeded up robust features (SURF), etc. (see Bay et al. (2008); Szeliski (2011) for details about interest points detection). The second step is to compute the descriptor around each interest point detected in the first step. Once again multiple techniques can be used for computing these descriptors such as the histogram of oriented gradients (HOG), wavelets, etc. (see Mohan et al. (2001); Dalal and Triggs (2005); Szeliski (2011) for details about computing the descriptors).

- **Quantization:** In quantization, three main steps are implemented. The first step is to determine the location of centroids of descriptors (also known as cluster centers) of each image and the second step is to assign each of the descriptor regions of the image to the closest cluster center, thereby grouping the descriptors into subsets known as visual words. Together these two steps are performed iteratively and constitute the process of clustering that can be performed by using k-means, fuzzy c-means, Gaussian mixture model, among others (see Bishop (2006) for more details about the clustering techniques). Finally, the third step of quantization is to compute the frequency, or histogram, of each one of the visual words. Thus, the result of the quantization process produces the histogram of the visual words of each image. It is worth noting that the quantization step involves random initialization of cluster centers that might affect the performance of the BoW modeling process.

- **Classification:** The classification is the last step of the BoW modeling process in which the decision rule, or boundary, is learned to distinguish between multiple classes. Many techniques can be used for the classification purpose such as decision tree classifier (DTC), linear discriminant classifier (LDC), support vector machine (SVM), among others. Similar to the quantization step, the classification step may involve random initialization of parameters that have a direct influence on the BoW modeling performance.

3. Preliminaries

Before detailing the proposed technique, we summarize several mathematical definitions and concepts that are helpful in developing the suggested ESRO-BoW. Readers familiar with the measure and probability theories may skip this section while those interested to know more about the preliminary concepts below may review them from related texts (e.g., Rudin (1987); Billingsley (1995)).

Suppose that we are given a set S and consider that Σ is a collection of subsets of S.

Definition 1: Σ is called a σ-algebra in S if the following conditions are met Rudin (1987).

1. S is a subset of Σ, i.e. $S \in \Sigma$.
2. If $A \in \Sigma$, then $\bar{A}$[1] $\in \Sigma$. Thus, all subsets of Σ should have their complements, relative to S, to be subsets of Σ as well. This property is called closed under the complement operation.
3. If $A = \bigcup_{n=1}^{\infty} A_n$ with $A_n \in \Sigma$ $(n = 1, 2, 3, \dots)$, then $A \in \Sigma$. This means that the countable union of subsets of Σ is a subset of Σ as well. This property is called closed under countable union.

If Σ is a σ-algebra in S, then S is called a measurable space that is usually denoted as a tuple (S, Σ). Let us denote $\mathcal{B}$ to be the smallest σ-algebra in S. $\mathcal{B}$ is called a Borel set of S if every open set of S is a subset of $\mathcal{B}$. Let $(\Omega, \mathcal{F}, p)$ be the probability space where Ω is a sample space[2], $\mathcal{F}$ is a set of events[3], and p is a probability measure[4](Billingsley, 1995). Now, a random variable X can be defined to be a mapping $X : \Omega \to S$ such that every Borel set $\mathcal{B} \in \Sigma$ has its pre-image in $\mathcal{F}$, i.e. $X^{-1}(B) \in \mathcal{F}$. This formal definition of random variable enables characterization of conditions, given below, for the concept of convergence of random variables. These concepts will be exploited in the subsequent sections.

Definition 2: Consider $\{X_n\}$ to be a sequence of random variables. $\{X_n\}$ is said to be converging in probability to X^* if for every $\varepsilon > 0$, we have (Billingsley, 1995; Grimmett and Stirzaker, 2001)

$$\lim_{n\to\infty} p(|X_n - X^*| > \varepsilon) = 0. \tag{1}$$

Establishing the convergence in probability of a random variable may be practically difficult. Thus, a more practical approach of assessing the convergence of a random variable, called convergence in mean, is provided in the definition below.

Definition 3: The random variable $\{X_n\}$ is said to be converging in the r-th absolute moment of $(X_n - X^*)$ if the expected value of $|X_n - X^*|^r$ tends to zero as n tends to ∞ (Billingsley, 1995), i.e.,

$$\lim_{n\to\infty} E(|X_n - X^*|^r) = 0, \tag{2}$$

where $E(\cdot)$ is the expected value.

Definition 3 provides a practical approach to check the convergence of a random variable since the convergence in mean implies the convergence in probability. However, in practice, there are situations where the random variable is a result of a certain computational process, as in this paper, and the convergence in mean may become computationally expensive and cumbersome to be validated. Thus, the characteristic of ergodicity of a stochastic process can simplify the determination of the convergence of a random variable from a certain subset of samples of the variable, as illustrated by the following definition of ergodicity.

Definition 4: A discrete random variable $X_n = \{x_1, x_2, ..., x_I\}$ is said to be an ergodic process in mean if $\hat{\mu}_{X_n}$, described by

$$\hat{\mu}_{X_n} = \frac{1}{I}\sum_{i=1}^{I} x_i, \tag{3}$$

converges to the ensemble average $E(X_n)$ as $I \to \infty$ (Papoulis and Pillai, 2002).

Thus, checking the ergodicity of the random variable is critical because if the random variable is an ergodic process, then the computational burden will be significantly reduced, as will be seen in the sequel. Throughout this paper, we use the notation $\mathcal{N}(\mu, \sigma^2)$ to denote a normal distribution with mean μ and standard deviation σ^2.

4. Enhanced Stochastically Robust and Optimized Bag-of-Words (ESRO-BoW) Modeling Process

The ESRO-BoW is developed to achieve two main objectives: guaranteeing the robustness against random initialization and estimating the optimal model order of the modeling process. Below is a detailed description of the stages through which the ESRO-BoW can be realized.

4.1. Robustness Against Random Initialization

Assume that the number of visual words M is given to model a certain set of images G. Let C be the performance achieved by using the BoW model $\mathcal{M}_{\mathrm{BoW}}$ in classifying the image set G. In order to have a confident evaluation of the BoW modeling process, we propose to perform the given task I times and we denote the resulting performance of the i^{th} iteration to be x_i. Then, one can form a random variable, denoted as X_1, representing the values of the performance for the I times as follows

$$X_1 = \{x_1, x_2, ..., x_I\}. \tag{4}$$

By repeating the above process N times, a sequence of random variables results that is denoted as X_{T} and given by

$$X_{\mathrm{T}} = \{X_1, X_2, ..., X_N\}. \tag{5}$$

The convergence in mean of the discrete random variable X_n, $n = 1, 2, \dots N$, to a random variable X^* implies the convergence in probability (Billingsley, 1995), i.e., it follows that X^* is the most probable random variable of X_{T}. Thus, the estimated X^* is the robust performance of the BoW model despite the existence of random initialization. Based on Definition 3 and using the definition of the limit, it follows that $\lim_{n\to\infty} E(|X_n - X^*|^r) = 0$ implies that for any real $\varepsilon \in R$ and $\varepsilon > 0$ there exists $N \in Z^+$,

[1] $\bar{A}$ is the complement of A.

[2] A sample space of an experiment is the set of all possible outcomes of the given experiment.

[3] An event is a set of outcomes including zero.

[4] A probability measure is a real-valued function assigning numbers to the events of the experiments.

Algorithm 1 Robust Bag-of-Words (R-BoW) Modeling

```
Inputs:
    Enter the set of images $G$
Initialize:
    A random variable $X \sim \mathcal{N}(\mu, \sigma^2)$, number of
    visual words $M$, tolerance $\varepsilon$, and positive
    integer $r$
$n \leftarrow 0$
repeat
    $n \leftarrow n + 1$
    for $i = 1$ to $I$ do
        Estimate the BoW model $\mathcal{M}_{\mathrm{BoW}}$ for classifying the set of images $G$
        Compute the performance $C$ of $\mathcal{M}_{\mathrm{BoW}}$
        $x_i \leftarrow C$
    end for
    $X_n \leftarrow \{x_1, ..., x_I\}$, $\mu_n \leftarrow E(X_n)$, $\sigma_n^2 \leftarrow E(X_n^2) - E(X_n)^2$
    $m_r \leftarrow E(|X_n - X|^r)$, $X \leftarrow X_n$
until $m_r < \varepsilon$
$\mu^* \leftarrow \mu_n$, $\sigma^* \leftarrow \sigma_n$, $X^* \leftarrow \mathcal{N}(\mu^*, \sigma^{*2})$
Outputs:
    $X^*$
```

where Z^+ is the set of positive integers, such that for every $n \in Z^+$ and $n \geq N$ we have (Rudin, 1976)

$$|E(|X_n - X^*|^r)| < \varepsilon. \tag{6}$$

Thus, if ε is chosen to be arbitrarily small, then one can use (6) to measure the convergence of the random variable $X_n, n = 1, 2, \ldots N$. Based on the notion above, Algorithm 1 is developed in which a Robust BoW (R-BoW) classifier is determined by employing (6) as a stopping criterion for the convergence of the performance, of the BoW model, that is considered as a sequence of random variables X_{T}.

The random variable X^* represents the limit to which the sequence X_n converges and its mean μ^* gives the expected value of the robust performance for the BoW modeling process despite the randomness of the initialization. As shown below, the expected value of the robust performance plays a key role in estimating the optimal performance that can be obtained using the proposed modeling technique. Algorithm 1 is applicable for both ergodic and non-ergodic processes. However, for the case of ergodic processes, which happens to be the case with many BoW modeling processes, one can simplify Algorithm 1 by relaxing the need for iterations of the outer loop since only one run of the loop is sufficient for ergodic processes. Thus for the case of ergodic BoW modeling processes, one can use Algorithm 2 to produce a Stochastically Robust-BoW (SR-BoW) modeling technique. The SR-BoW modeling technique is applicable when the modeling process is ergodic resulting in a significantly faster implementation compared to the R-BoW implementation of Algorithm 1. Though Algorithms 1 and 2 address the sensitivity of the BoW modeling process with respect to random initialization, the selection of the optimal model order, i.e., the number of visual words, remains open. Thus, the next subsection addresses the estimation of the number of words resulting in the optimal performance.

Algorithm 2 Stochastically Robust Bag-of-Words (SR-BoW) Modeling

```
Inputs:
    Enter the set of images $G$
Initialize:
    Number of visual words $M$ and a positive
    integer $r$
for $i = 1$ to $I$ do
    Estimate the BoW model $\mathcal{M}_{\mathrm{BoW}}$ for classifying the set of images $G$
    Compute the performance $C$ of $\mathcal{M}_{\mathrm{BoW}}$
    $x_i \leftarrow C$
end for
$X \leftarrow \{x_1, ..., x_I\}$, $\mu^* \leftarrow E(X)$, $\sigma^{*2} \leftarrow E(X^2) - E(X)^2$
$X^* \leftarrow \mathcal{N}(\mu^*, \sigma^{*2})$
Outputs:
    $X^*$
```

4.2. Order Selection of BoW Models

Suppose that the number of words M can take any value in the set $M_\Omega = \{2, 3, ..., M_u\}$. Consider that $X_n \sim p(X_n|\theta)$ and let $\theta \sim p(\theta|M)$. Using the Bayes rule, we can obtain the posterior distribution of the parameter θ to be

$$p(\theta|X_n, M) = \frac{p(X_n|\theta, M)p(\theta|M)}{p(X_n|M)}. \tag{7}$$

Assuming the term $p(X_n|\theta, M)$ to be a Markov process, we obtain

$$p(X_n|\theta, M) = p(X_n|\theta). \tag{8}$$

Using (7) and (8) yields

$$p(\theta|X_n, M) = \frac{p(X_n|\theta)p(\theta|M)}{p(X_n|M)}. \tag{9}$$

From (9), we have

$$p(\theta|X_n, M) \propto p(X_n|\theta)p(\theta|M). \tag{10}$$

The posterior predictive distribution at the limit of the sequence X_n, i.e., X^*, is given by

$$p(X^*|X_n, M) = \int_\theta p(X^*|\theta)p(\theta|X_n, M)d\theta. \tag{11}$$

Let M^* be the optimal value of M that maximizes the posterior predictive distribution $p(X^*|X_n, M)$. From (10) and (11), it is obvious that $p(X^*|X_n, M)$ increases as $p(X^*|\theta)$ and $p(\theta|M)$ increase. Let $X^*(M)$ denote the limit of the sequence of the random variables X_n for a given number of visual words M. Then, M^* is determined from

$$M^* = \arg\max_M E(X^*(M)). \tag{12}$$

Hence, (12) is a good measure for finding the optimal number of visual words M in the BoW modeling process. However, M is required to have known upper and lower bounds, i.e., $M \in [\underline{M}, \overline{M}]$ where $\underline{M}, \overline{M} \in Z^+$ and $\overline{M} > \underline{M}$. Thus, one can integrate (12) with the SR-BoW modeling technique to develop

BoW models that are simultaneously robust against random initialization and estimate the optimized model order. The resulting algorithm is a Stochastically Robust and Optimized-BoW (SRO-BoW) modeling technique that is detailed in Algorithm 3 above.

Algorithm 3 Stochastically Robust and Optimized Bag-of-Words (SRO-BoW) Modeling

```
Inputs:
    Enter the set of images $G$
Initialize:
    The step size $h$ and a positive integer $r$
$E_{M^*} \leftarrow 0$
for $M = \underline{M}$ to $\overline{M}$ do
    for $i = 1$ to $I$ do
        Estimate the BoW model $\mathcal{M}_{\text{BoW}}$ for classifying the set of images $G$
        Compute the performance $C$ of $\mathcal{M}_{\text{BoW}}$
        $x_i \leftarrow C$
    end for
    $X^*_M \leftarrow \{x_1, ..., x_I\}, \mu_M \leftarrow E(X^*_M)$
    $X \leftarrow X^*_M, \sigma^2_M \leftarrow E({X^*_M}^2) - E(X^*_M)^2$
end for
$M^* \leftarrow \arg\max_M \mu_M, X^*_{M^*} \leftarrow \mathcal{N}(\mu_{M^*}, \sigma^2_{M^*})$
Outputs:
    $X^*_{M^*}, M^*$
```

Though finding the optimal number of the visual words using Algorithm 3 is efficient, it requires a significant amount of computational cost. The main reason behind such computational burden in Algorithm 3 is a consequence of computing the expectation for all values of $M = \{Z^+ : Z^+ \in [\underline{M}, \overline{M}]\}$ that makes the algorithm complexity to be $O((\overline{M} - \underline{M}) \times I)$. To overcome this computational burden, we can assume that the expectation of $X^*(M)$ is locally maximum for zero difference with respect to M. That is,

$$\lim_{M \to M^*} \Delta E(X^*(M)) = 0. \tag{13}$$

Since M is an integer, it follows that $\Delta E(X^*(M))$ can be represented by $E(X^*(M)) - E(X^*(M - h))$ where h is the step size of the increment in M such that $h \in Z^+$ and $h \neq 0$. Thus, we obtain

$$\lim_{M \to M^*} E(X^*(M)) - E(X^*(M - h)) = 0. \tag{14}$$

From (14) and by invoking the definition of limits, for each real number $\varepsilon_M \in R$ and $\varepsilon_M > 0$ there exists a positive number δ such that $|M - M^*| < \delta$ implies

$$|E(X^*(M)) - E(X^*(M - h))| < \varepsilon_M. \tag{15}$$

Based on (15), an Enhanced Stochastically Robust and Optimized-BoW (ESRO-BoW) modeling technique is developed and is detailed in Algorithm 4 wherein (15) is used as a stopping criterion for finding the optimal number of visual words that optimizes the performance of the BoW models. The convergence of the difference in expectation $E_{M^*} - E_{M^o}$ to the open ball $B_{\varepsilon_M}(E(X(M^*)))$ becomes feasible before the outer loop ends rendering the complexity to be less than $O((\overline{M} - \underline{M}) \times I)$. Figure 1 visualizes the convergence of the ESRO-BoW modeling process where each row represents a random variable of the classification performance for I-times execution of the BoW for each value of M. It is obvious from Figure 1 that the stopping condition (15) is checked with each increment of M. As soon as (15) is met, the algorithm stops otherwise the value of M is increased, the random variable of the classification performance is re-collected, and the computations are re-iterated until (15) is met.

Algorithm 4 Enhanced Stochastically Robust and Optimized Bag-of-Words (ESRO-BoW) Modeling

```
Inputs:
    Enter the set of images $G$
Initialize:
    A random variable $X \sim \mathcal{N}(\mu, \sigma^2)$, step size $h$,
    tolerance $\varepsilon_M$ and positive integer $r$
$M \leftarrow \underline{M} - h, E_{M^*} \leftarrow 0$
repeat
    $M \leftarrow M + h$
    for $i = 1$ to $I$ do
        Estimate the BoW model $\mathcal{M}_{\text{BoW}}$ for classifying the set of images $G$
        Compute the performance $C$ of $\mathcal{M}_{\text{BoW}}$
        $x_i \leftarrow C$
    end for
    $X^*_M \leftarrow \{x_1, ..., x_I\}, \mu_M \leftarrow E(X^*_M)$
    $\sigma^2_M \leftarrow E({X^*_M}^2) - E(X^*_M)^2$
    if $\mu_M > E_{M^*}$ then
        $E_{M^o} \leftarrow E_{M^*}, E_{M^*} \leftarrow \mu_M, \mu^*_M \leftarrow \mu_M, {\sigma_{M^*}}^2 \leftarrow \sigma^2$
    end if
until $|E_{M^*} - E_{M^o}| < \varepsilon_M$
$M^* \leftarrow M, X^*_{M^*} \leftarrow \mathcal{N}(\mu^*, {\sigma^*}^2)$
Outputs:
    $X^*_{M^*}, M^*$
```

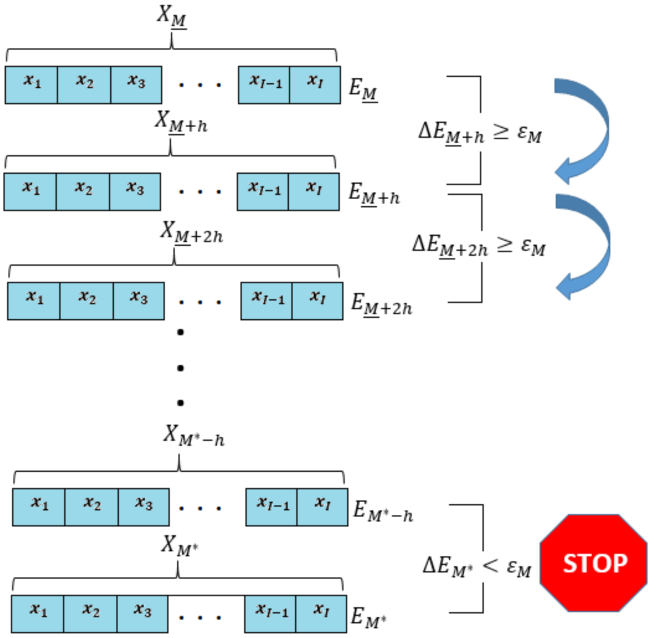


Fig. 1. Schematic diagram of the convergence of the ESRO-BoW modeling process.

5. Experimental Validation

To evaluate the performance of the ESRO-BoW, we consider the image classification problem of the Caltech 101 image set composed of 101 categories with 40 – 800 images per category (see Fei-Fei et al. (2004) for more details about this image set). The ESRO-BoW modeling technique of Algorithm 4 was employed in classifying this standard image set. The value of ε_M was taken to be 0.2%, r was considered to be 1, I was taken to be 1,000 samples and X was initialized to be $\mathcal{N}(0, 0.1)$. The grid method was used to select the locations of feature points while SURF method was employed to extract the features from the located feature points. The k-means clustering technique was employed in the quantization step and Quadratic Support Vector Machine (QSVM) classification technique was considered in the classification step.

To assess the performance of the ESRO-BoW of Algorithm 4, the aforementioned image set was employed for three cases: 5, 15, and 30 training images. Figure 2 (a), (b), and (c) show that the distribution of the ESRO-BoW classification performance for the cases of 5, 15, and 30 training images, respectively, where the axes of the graphs represent the number of words, classification performance, and density of the classification performance. Even though the ESRO-BoW algorithm stopped at $M = 900$, $M = 1,200$, and $M = 1,700$ for the cases of 5, 15, and 30 training images, respectively, we extended the graphs in Figure 2 by computing the corresponding values until 2,000 words to gain a full understanding of the performance of the ESRO-BoW modeling process. From Figure 2 (a), (b), and (c), it can be seen that the classification performance in all three cases produces a stochastic behavior in the sense that their performance is not a fixed value for a fixed number of words. Such a stochastic behavior stems from the nature of BoW and its dependence on the arbitrary initialization of sets of parameters, in the quantization and classification steps, rendering the performance to vary with multiple executions of the modeling process for a fixed value of M.

To examine the performance of the three ESRO-BoW executions shown in Figure 2 (a), (b), and (c), we plotted the expectation of the obtained distributions in Figure 3 (a) that concretely illustrates the performance for each of the considered ESRO-BoW implementation. The local maximum expected value was realized with $M = 900$, $M = 1,200$, and $M = 1,700$ words for the cases of 5, 15, and 30 training images, respectively, where the ESRO-BoW modeling process stopped. Thus, in addition to its robustness, the ESRO-BoW technique can identify the number of words resulting in local maximum expected value of the classification performance. The local maximum expected value of the classification performance was found to be 52.26%, 70.01%, and 75.02% for the cases of 5, 15, and 30 training images, respectively. If we examine the behavior of the expected value of the classification performance for the three ESRO-BoW implementations shown in Figure 3 (a), one can see that the increment of M beyond the aforementioned points of local maxima does not necessarily imply the increment in the classification performance. In fact, excessively large values of M can lead to overfitting that degrades the performance of the modeling process. Figure 3 (b) shows the standard deviation of the considered ESRO-BoW technique with respect to the number of visual words for the cases of 5, 15, and 30 training images. It is clear from Figure 3 (b) that the value of the standard deviation at 900 words, for the case of 5 training images, is more than its neighborhood which gives the impression that along with the highest expected value of the classification performance shown in Figure 3 (a), we obtain a local maximum classification performance when the number of words is 900. Likewise for the 15, and 30 training images, the values of their standard deviation are higher than their corresponding neighborhoods implying the fact that local maxima of the classification performance are obtained at 1,200 and 1,700 words for the cases of 15 and 30 training images, respectively. Fixing the number of words to be 900, 1,200, and 1,700 for the cases of 5, 15, and 30 training images, respectively, we graphed the histograms of the 1,000 samples of each case of the considered number of training images as shown in Figure 4. The value of the ensemble average of the 1,000 samples of the performance of the ESRO-BoW was computed to be 53.15%, 71.98%, and 76.74%, respectively. The error between the ensemble averages and the means were computed to be 0.89%, 1.97%, and 1.72% with 5, 15, and 30 training images, respectively. It is obvious that the values of errors between the ensemble average and mean are relatively small which, according to Definition 4, suggests that the given process is an ergodic process. Note that the aforementioned errors result because we considered a random variable of only 1,000 samples to model the process. Our experiment shows that as the number of samples is increased from 1,000, the error between ensemble averages and the means continues to reduce. However, for computational efficiency, we have opted to limit the number of samples to 1,000.

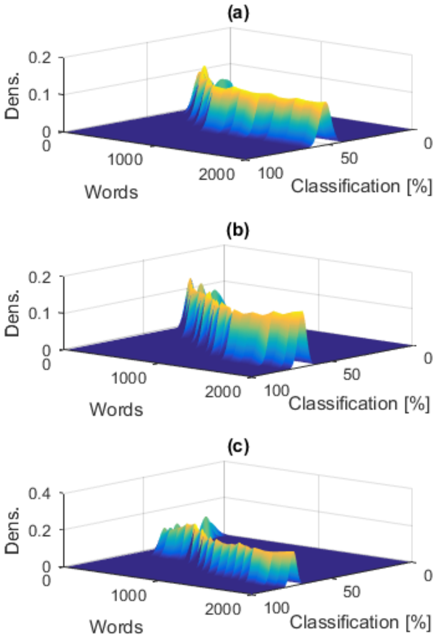


Fig. 2. ESRO-BoW 3-dimensional classification performance with Caltech 101 image set: (a) 5 training images, (b) 15 training images, and (c) 30 training images.

The maximum classification performance obtained using ESRO-BoW for the cases of 5 (with 900 words), 15 (for 1,200

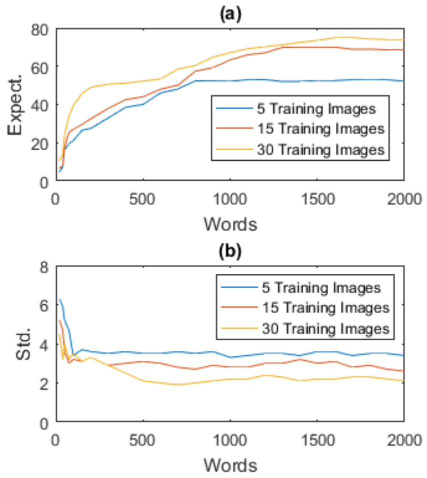


Fig. 3. Caltech 101 **image set performance for** 5 **training images,** 15 **training images, and** 30 **training images: (a) The expected value of the classification performance, (b) The standard deviation of the classification performance.**

words), and 30 (with 1, 700 words) training images are 67.61%, 80.52%, and 85.35%, respectively, that represent local maxima for the considered numbers of training images. Table 1 summarizes the classification performance of the ESRO-BoW modeling technique compared with that of alternative, well-known, state-of-the-art works used for the classification of the Caltech 101 image set. The results of Table 1 lead us to conclude that the ESRO-BoW modeling technique yields a significant enhancement when employed for classifying the Caltech 101 image set. The main reasons behind such improvement with the ESRO-BoW modeling are the use of the optimal number of words and accommodating the problem of random initialization encountered in the steps of the BoW modeling process.

Non-optimal performance resulting from random initialization and non-optimal parameters cause performance degradation due to being trapped at local minima in the quantization and classification steps of the BoW process since both involve differentiations in their implementation. It is worth noting that the BoW modeling performance can be enhanced by using the k-means++ clustering technique in the quantization step. However, employing the k-means++ resulted in classification performance of 61.03%, 75.74%, and 78.29% for the cases of 5, 15, and 30 respectively which is less than the corresponding performance of the ESRO-BoW modeling technique, giving an impression that the k-mean++ does not fully address the problems of random initialization that are encountered in the quantization and classification steps of BoW process.

The model order, or the number of words M, of the ESRO-BoW is tuned such that the model order resulting in the highest classification performance is selected for each one of the aforementioned cases. Employing (15) as a stopping condition is the main reason behind finding the optimal value of M since it relies on finding the value that makes the difference equation of the expected value, with respect to M, to be zero (see (13)-(15)). Even though such difference equations can not be claimed to result in a global maximum of the expected value, it can produce the value of M that locally maximizes the expected value of the classification performance. Thus, employing the ESRO-BoW

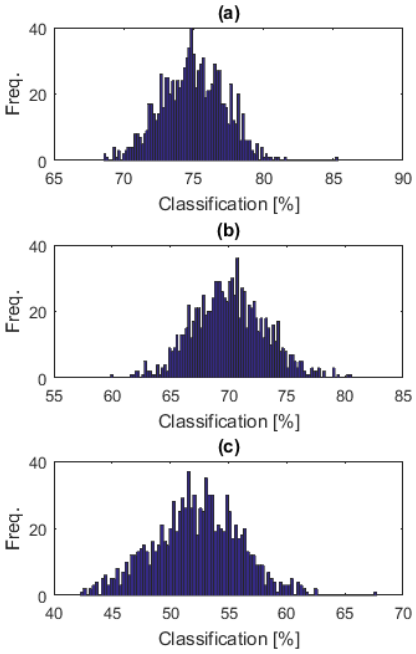


Fig. 4. Histograms of the performance of ESRO-BoW technique applied to Caltech101 **image set: (a)** 5 **training images with** $M = 900$**, (b)** 15 **training images with** $M = 1200$**, and (c)** 30 **training images with** $M = 1700$**.**

Table 1. Classification performance of Caltech101 image set for 5, 15, **and** 30 **training images situations.**

Training Images	**5**	**15**	**30**
Method	**Performance**		
SPM-BoW (Lazebnik et al., 2006)		56.4	64.6
SVM-KNN (Zhang et al., 2006)	46.6	59.05	66.23
NBNN (Boiman et al., 2008)	44.2	59.05	66.23
NBNN Kernel (Tuytelaars et al., 2011)		61.3	69.6
LSPM (Yang et al., 2009)		67.0	73.2
M-HMP (Bo et al., 2013)			82.5
GPP (Xie et al., 2014)	61.90	76.03	82.45
LLC+SDP (Mu et al., 2015)	53.6	69.7	77.1
ESRO-BoW	67.61	80.52	85.35

technique relaxes the need to know the number of words and it automatically computes the value of M to maximize the classification performance. During the execution of the ESRO-BoW modeling technique, the step size h was increased with multiple, arbitrarily chosen values such that it can realize the ESRO-BoW in a reasonable time. However, optimizing the value of h while considering the computational time as a cost function can result in an optimal value of h.

6. Conclusion

In this paper, an ESRO-BoW modeling technique is developed to address the problems of sensitivity of the BoW performance to random initialization and number of visual words. The robustness against parameter initialization is realized by considering the performance as a sequence of random variables and the limit of the resulting sequence is estimated using the convergence in mean. The sequence of obtained random variables results in a distribution of performance which helps determine the optimal number of words that maximizes the expected value of the limit of the random variable. Thus, the ESRO-

BoW technique achieves simultaneous robustness against random parameter initialization and optimization of the number of words. Experimental validation was conducted for classifying the Caltech 101 image set and application of the ESRO-BoW technique was shown to yield a significantly superior classification performance relative to several state-of-the-art techniques. Despite the excellent performance reported for the ESRO-BoW, the step size of the increment of the number of words was altered in an *ad hoc* fashion which may adversely affect the computational cost. Thus, future research will focus on developing a formal methodology to determine the step size increment that minimizes the computational time of the ESRO-BoW technique. Moreover, deriving a closed form relationship that maps the number of words with the shape of the classification performance will give a direct estimation of the optimal number of visual words.

Acknowledgments

This work is partially supported by NSF grants DRK-12 DRL: 1417769, ITEST DRL: 1614085, and RET Site EEC: 1542286. The authors would like to thank the anonymous reviewers for their constructive feedbacks that significantly enhanced the quality of the paper.

References

Bay, H., Ess, A., Tuytelaars, T., Gool, L.V., 2008. Speeded-up robust features (SURF). Comp. Vision and Image Understanding 110, 346–359.

Billingsley, P., 1995. Probability and Measure, 3rd Ed. John Wiley and Sons, Inc., New York, NY, USA.

Bishop, C.M., 2006. Pattern Recognition and Machine Learning. Springer, New York, USA.

Bo, L., Ren, X., Fox, D., 2013. Multipath sparse coding using hierarchical matching pursuit, in: IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR. pp. 660–667.

Boiman, O., Shechtman, E., Irani, M., 2008. In defense of nearest-neighbor based image classification, in: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8.

Bolovinou, A., Pratikakis, I., Perantonis, S., 2013. Bag of spatio-visual words for context inference in scene classification. Pattern Recognition 46, 1039–1053.

Burghouts, G.J., Schutte, K., 2013. Spatio-temporal layout of human actions for improved bag-of-words action detection. Pattern Recognition Letters 34, 1861–1869.

Camilleri, D., Prescott, T., 2017. Analysing the limitations of deep learning for developmental robotics, in: Mangan, M., Cutkosky, M., Mura, A., Verschure, P.F., Prescott, T., Lepora, N. (Eds.), Biomimetic and Biohybrid Systems, Springer International Publishing. pp. 86–94.

Csurka, G., Dance, C.R., Fan, L., Willamowski, J., Bray, C., 2004. Visual categorization with bags of keypoints, in: Workshop on Statistical Learning in Computer Vision, ECCV, Prague, Czech Republic. pp. 1–22.

Dalal, N., Triggs, B., 2005. Histograms of oriented gradients for human detection, in: IEEE Comp. Soc. Conf. Comp. Vision and Pattern Recog., pp. 886–893.

Fei-Fei, L., Fergus, R., Perona, P., 2004. Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories, in: Conference on Computer Vision and Pattern Recognition Workshop, pp. 178–178.

Fu, Z., Lu, G., Ting, K.M., Zhang, D., 2011. Music classification via bag-of-features approach. Pattern Recognition Letters 32, 1768–1777.

Grimmett, G., Stirzaker, D., 2001. Probability and Random Process, 3rd Ed. Oxford University Press, Inc., New York, NY, USA.

Hernández-Vela, A., Bautista, M.A., Perez-Sala, X., Ponce-López, V., Escalera, S., Baró, X., Pujol, O., Angulo, C., 2014. Probability-based dynamic time warping and bag-of-visual-and-depth-words for human gesture recognition in rgb-d. Pattern Recognition Letters 50, 112–121.

Iosifidis, A., Tefas, A., Pitas, I., 2014. Discriminant bag of words based representation for human action recognition. Pattern Recognition Letters 49, 185–192.

Karakasis, E., Amanatiadis, A., A.Gasteratos, Chatzichristofis, S., 2015. Image moment invariants as local features for content based image retrieval using the bag-of-visual-words model. Pattern Recognition Letters 55, 22–27.

Lazebnik, S., Schmid, C., Ponce, J., 2006. Beyond bags of features: spatial pyramid matching for recognizing natural scene categories, in: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, New York, NY. pp. 1–8.

Lecun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444.

Mohan, A., Papageorgiou, C., Poggio, T., 2001. Example based object detection in images by components. IEEE Trans. Pattern Anal. and Machine Intell. 23, 349–361.

Mu, G., Liu, Y., Wang, L., 2015. Considering the spatial layout information of bag of features (bof) framework for image classification. PLOS ONE 10, 1–13.

Nazir, S., Yousaf, M., Nebel, J., Velastin, S., 2018. A bag of expression framework for imrpoved human action recognition. Pattern Recognition Letters 103, 39–45.

Nguyen, A., Yosinski, J., Clune, J., 2015. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 427–436.

Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z.B., Swami, A., 2016. The limitations of deep learning in adversarial settings, in: IEEE European Symposium on Security and Privacy (EuroS P), pp. 372–387.

Papoulis, A., Pillai, S.U., 2002. Probability, Random Variables, and Stochastic Processes, 4th Ed. McGraw-Hill, Inc., New York, NY, USA.

Passalis, N., Tefas, A., 2017. Neural bag-of-features learning. Pattern Recognition 64, 277 – 294.

Rudin, W., 1976. Principles of Mathematical Analysis, 3rd Ed. McGraw-Hill, Inc., New York, NY, USA.

Rudin, W., 1987. Real and Complex Analyasis, 3rd Ed. McGraw-Hill, Inc., New York, NY, USA.

dos Santos, J., de Moura, E., da Silva, A.S., Cavalcanti, J., da Silva Torres, R., Vidal, M., 2015. A signature-based bag of visual words method for image indexing and search. Pattern Recognition Letters 65, 1–7.

Shen, L., Lin, J., Wu, S., Yu, S., 2014. Hep-2 image classification using intensity order pooling based features and bag of words. Pattern Recognition 47, 2419–2427.

Szeliski, R., 2011. Computer Vision: Algorithms and Applications. Springer-Verlag, London, UK.

Tuytelaars, T., Fritz, M., Saenko, K., Darrell, T., 2011. The NBNN kernel, in: International Conference on Computer Vision, Barcelona, Spain. pp. 1824–1831.

Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., Gong, Y., 2010. Locality-constrained linear coding for image classification, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 3360–3367.

Wu, S., Wong, H., 2012. Joint segmentation of collectively moving objects using a bag-of-words model and level set evolution. Pattern Recognition 45, 3389–3401.

Xie, L., Tian, Q., Wang, M., Zhang, B., 2014. Spatial pooling of heterogeneous features for image classification. IEEE Transactions on Image Processing 23, 1994–2008.

Yang, J., Jiang, Y.G., Hauptmann, A., Ngo, C.W., 2007. Evaluating bag-of-visual-words representations in scene classification, in: Proc. Int. Work. Mult. Inform. Retr., Augsburg, Germany. pp. 197–206.

Yang, J., Yu, K., Gong, Y., Huang, T., 2009. Linear spatial pyramid matching using sparse coding for image classification, in: IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL. pp. 1794–1801.

Zagoris, K., Pratikakis, I., Antonacopoulos, A., Gatos, B., Papamarkos, N., 2014. Distinction between handwritten and machine-printed text based on the bag of visual words model. Pattern Recognition 47, 1051–1062.

Zhang, H., Berg, A.C., Maire, M., Malik, J., 2006. SVM-KNN: discriminative nearest neighbor classification for visual category recognition, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), New York, USA. pp. 2126–2136.