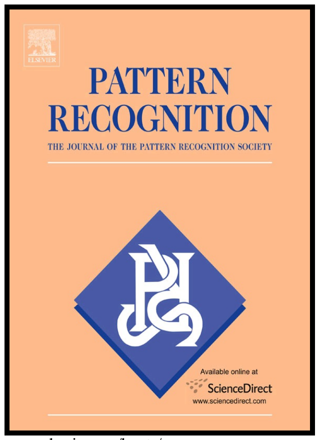
# Author's Accepted Manuscript




www.elsevier.com/locate/pr

Neural Bag-of-Features Learning

Nikolaos Passalis, Anastasios Tefas

PII: S0031-3203(16)30371-5
DOI: http://dx.doi.org/10.1016/j.patcog.2016.11.014
Reference: PR5957

To appear in: *Pattern Recognition*

Received date: 23 March 2016
Revised date: 18 August 2016
Accepted date: 16 November 2016

Cite this article as: Nikolaos Passalis and Anastasios Tefas, Neural Bag-of-Features Learning, *Pattern Recognitior* http://dx.doi.org/10.1016/j.patcog.2016.11.014

# Neural Bag-of-Features Learning

Nikolaos Passalis, Anastasios Tefas

*Department of Informatics, Aristotle University of Thessaloniki*
*Thessaloniki 54124, Greece Tel,Fax: +30-2310996304*

**Abstract**

In this paper, a neural learning architecture for the well-known Bag-of-Features (BoF) model, called Neural Bag-of-Features, is proposed. The Neural BoF model is formulated in two neural layers: a Radial Basis Function (RBF) layer and an accumulation layer. The ability of the Neural BoF model to improve the classification performance is demonstrated using four datasets, including a large-scale dataset, and five different feature types. The gains are two-fold: the classification accuracy increases and, at the same time, smaller networks can be used, reducing the required training and testing time. Furthermore, the Neural BoF natively supports training and classifying from feature streams. This allows the proposed method to efficiently scale to large datasets. The streaming process can be also used to introduce noise and reduce the over-fitting of the network. Finally, the Neural BoF provides a framework that can model and extend the dictionary learning methodology.

*Keywords:* Bag-of-Features, RBF Neural Networks, Dictionary Learning



## 1. Introduction

Many machine learning problems involve dealing with objects that are represented as a collection of *feature vectors*. For example, the typical pipeline of an image recognition system is composed of the following steps:

1. First, multiple feature vectors, such as SIFT descriptors [1], HOG descriptors [2], or even raw pixel patches [3], are extracted from an image,
2. then, these feature vectors are used to create a constant length vector representation of the image, and
3. finally, this vector is fed to a classifier that recognizes the content of the image.

Usually, a large number of feature vectors are extracted from each image. Furthermore, the number of the extracted features vectors might vary, especially when interest point detectors are used [1], [4]. Therefore, it is critical to "aggregate" the extracted vectors into a meaningful constant length representation in order to efficiently measure the similarity of two images.

This problem can be formally defined as follows: Let $\mathcal{X} = \{x_i\}_{i=1}^{N}$ be a set of $N$ objects. Each object $x_i$ consists of $N_i$ feature vectors: $\mathbf{x}_{ij} \in \mathbb{R}^D$ ($j = 1...N_i$), where $D$ is the dimensionality of the extracted features. Aggregate the

*Email addresses:* passalis@csd.auth.gr (Nikolaos Passalis), tefas@aiia.csd.auth.gr (Anastasios Tefas)