FEATURE NORMALIZATION FOR PART-BASED IMAGE CLASSIFICATION

Lingxi Xie\textsuperscript{1}, Qi Tian\textsuperscript{2}, and Bo Zhang\textsuperscript{1}

\textsuperscript{1}State Key Laboratory of Intelligent Technology and Systems (LITS)  
\textsuperscript{1}Tsinghua National Laboratory for Information Science and Technology (TNList)  
\textsuperscript{1}Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China  
\textsuperscript{2}Department of Computer Science, University of Texas at San Antonio, Texas, USA  
\textsuperscript{1}198808xc@gmail.com,\textsuperscript{2}qitian@cs.utsa.edu,\textsuperscript{3}dcszb@mail.tsinghua.edu.cn

ABSTRACT

Part-based Bag-of-Features (BoF) models such as Spatial Pyramid Matching (SPM) play an important role in image classification. Before sending the feature vectors into classifiers for training and testing, it is required to normalize them in order to approximately equalize ranges of the attributes and make them have comparable effects in distance computation. Although some works have been focused on general feature normalization, we do not see any discussion on specialized normalization algorithms for part-based BoF models.

In this paper, we fill in the blank with extensive experiments and discussions. Based on solid normalization parameters (power and coefficient), we further study two straightforward part-based properties, i.e., the independent assumption and the hierarchical-contribution assumption, to scale the feature super-vectors separately. Finally, we test our algorithm on challenging image sets, i.e., Caltech101 and CUB-200-2011, for general and fine-grained classification, and show its efficiency, scalability and adaptability in both scenarios.

Index Terms— Part-Based Bag-of-Features Models, Image Classification, Feature Normalization, Experiments.

1. INTRODUCTION

Large-scale image classification has been a hot topic for many years. It is a challenge towards image understanding and implies a wide range of real applications. Today, one of the most popular methods of image classification is to represent images with long super-vectors, and use a generalized classifier (such as SVM [1]) for training and testing.

The Bag-of-Features (BoF) model [2] [3] is widely used for image representation. It is a statistics-based model which summarizes local features in a sparse vector. The major shortcomings of the BoF model come from the well-known gap between low-level pixels and high-level concepts [4] [5]. In recent years, new modules were proposed to bridge the semantic gap and provide more robust image representations. System with the state-of-the-art techniques [6] produces relatively discriminative representation in terms of super-vectors.

Before using the super-vectors for classification, feature normalization is considered as an important data preprocessing step to avoid attributes in greater numeric ranges dominating those in smaller numeric ranges. Despite the existing works on feature normalization applied to various classification models, there still lacks a specialized algorithm for part-based BoF models.

In this respect, we conduct an in-depth study on part-based feature normalization by proposing several novel algorithms and comparing them with baseline systems. We claim a two-fold contribution. First, we formulate power and coefficient, two normalization parameters, and show their tremendous impact on classification accuracies in experiments. Second, taking the part-based properties into consideration, we develop specialized normalization approaches for part-based BoF models. Integrating both techniques obtains an enhanced algorithm to outperform the state-of-the-art models.

The rest of this paper is organized as follows. First, we provide a brief overview of the BoF model in Section 2. Then Section 3 proposes several novel algorithms for feature normalization. After extensive experiments and discussions are given in Section 4, we draw our conclusions in Section 5.

2. THE BAG-OF-FEATURES MODEL

The Bag-of-Features model starts from a raw image $I$:

$$I = (a_{ij})_{W \times H}$$  \hspace{1cm} (1)

where $a_{ij}$ is the pixel at position $(i, j)$. For better local representation, we extract a set of SIFT [7] descriptors $D$:

$$D = \{d_1, d_2, \ldots, d_M\}$$  \hspace{1cm} (2)

where $d_m$ denotes the description vector of the $m$-th descriptor. $M$ is the total number of descriptors.

After descriptors have been extracted, they are quantized to be compact. For this purpose, we train a codebook $C$, which is a $B \times D$ matrix consisting of $B$ vectors with dimension $D$, each of which is called a codeword. The codebook size $B$, is thousands in classification tasks.
The individually pooled representation vectors are finally concatenated as a super-vector:

$$F = (f_1, f_2, \ldots, f_S)$$  \hspace{1cm} (6)

which is the output of the part-based BoF model.

### 3. FEATURE NORMALIZATION

Feature normalization, or feature scaling, is a basic technique for data pre-processing. With a clear motivation to approximately equalize the range and weight of input attributes, researchers proposed various techniques to apply on different classification models, such as Support Vector Machines [11] [12] [13], Naive Bayes Classifier [14], Hidden Markov Model Estimation [15], Kernel Fisher Discriminant Analysis [16], and even image retrieval systems [17].

#### 3.1. Normalization Power and Coefficient

One of the simplest and most widely-used formulation is the $\ell_p$-normalization, which calculates the super-vector’s linear projection on the $\ell_p$-norm unit hyper-sphere:

$$\tilde{F} = \frac{F}{\|F\|_p}$$  \hspace{1cm} (7)

where $\|F\|_p$ is the $\ell_p$-norm: $\|F\|_p = (\sum f_i^p)^{1/p}$, and $p$ is named the normalization power. In most cases, we use SVM for image classification. Since the formulation of SVM is quite sensitive to the numerical ranges of input data, it is reasonable to choose a proper normalization coefficient for robust feature super-vectors. For this, we modify (7) as:

$$\tilde{F} = w \times \frac{F}{\|F\|_p}$$  \hspace{1cm} (8)

where $w$ is the normalization coefficient.

We plot the classification results using (8) in Figure 2. Detailed settings are mentioned in Section 4. It is shown that proper $p$ and $w$ are very important for normalization. Further, we claim the ability of $\ell_1$-normalization by showing its comparable performances with $\ell_2$ and $\ell_\infty$, under a large coefficients $w = 100$ which is suitable for all powers. This refutes the opinion that $\ell_1$-normalization would cause classification accuracies drop dramatically [8]. In later experiments, we fix $w = 100$ for all normalization powers.

#### 3.2. Separate Normalization by Parts

Till now, we make no efforts on designing specialized normalization methods for part-based BoF models. If we assume that each part contributes independently to the entire model (the independent assumption), it is straightforward to keep...
the same \( \ell_p \)-norm individually in all the pooling bins, which leads in normalizing \( f_1, f_2, \ldots, f_S \) in (6) separately:

\[
\tilde{f}_s = \frac{w}{S^{1/p}} \frac{f_s}{\|f_s\|_p}, \quad s = 1, 2, \ldots, S
\]  

(9)

and concatenating \( \tilde{f}_1, \tilde{f}_2, \ldots, \tilde{f}_S \) again:

\[
\tilde{F} = (\tilde{f}_1, \tilde{f}_2, \ldots, \tilde{f}_S)
\]  

(10)

The modified normalization coefficient \( w S^{1/p} \) in (9) aims to preserve \( \|F\|_p = w \) as it is in (8).

### 3.3. Hierarchical Weighting Scheme (HRC)

It is not always true that each part contributes equally. Most often, high-level pooling bins consist of more basic regions and therefore are more robust and discriminative. Denote \( k_s \) as the number of basic regions contained in \( s \)-th pooling bin, and assume that \( s \)-th bin’s contribution is proportional to \( k_s^{1/p} \) (the hierarchical-contribution assumption). Appending the fixed constraint \( \|F\|_p = w \) obtains the equations system:

\[
\left\{ \begin{align*}
    w_p^s & \propto k_s \\
    \sum_{s=1}^S w_p^s &= w^p
\end{align*} \right. 
\]  

(11)

Solving (11) gives a group of new coefficients for part-wise normalization, in which we enhance the spatial weights on high-level pooling bins to emphasize global information.

## 4. EXPERIMENTAL RESULTS

This section shows experimental results on different part-based BoF models: SPM \([9]\) \([6]\) on Caltech101 \([18]\), and

### 4.1. The Caltech101 Dataset

The Caltech101 Dataset \([18]\) is a basic object collection containing 9144 images from 102 categories. Sample images are
Table 2. Classification results on CUB-200-2011 using different models and normalization techniques. The normalization coefficient $\nu$ is set as 100 in all cases. The standard deviation of each result is around 0.3%.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>LLC</th>
<th>LLC-HP</th>
<th>GPP-HP</th>
</tr>
</thead>
<tbody>
<tr>
<td>No normalization</td>
<td>25.58</td>
<td>27.55</td>
<td>30.67</td>
</tr>
<tr>
<td>Global $\ell_1$</td>
<td>27.61</td>
<td>29.85</td>
<td>33.22</td>
</tr>
<tr>
<td>Global $\ell_2$</td>
<td>27.06</td>
<td>29.30</td>
<td>32.96</td>
</tr>
<tr>
<td>Global $\ell_\infty$</td>
<td>25.04</td>
<td>27.41</td>
<td>30.71</td>
</tr>
<tr>
<td>Separate $\ell_1$</td>
<td>24.58</td>
<td>26.94</td>
<td>31.84</td>
</tr>
<tr>
<td>Separate $\ell_2$</td>
<td>30.93</td>
<td>32.75</td>
<td>35.98</td>
</tr>
<tr>
<td>Separate $\ell_\infty$</td>
<td>27.73</td>
<td>29.67</td>
<td>31.92</td>
</tr>
<tr>
<td>Separate $\ell_1$ + HRC</td>
<td>—</td>
<td>28.12</td>
<td>33.85</td>
</tr>
<tr>
<td>Separate $\ell_2$ + HRC</td>
<td>—</td>
<td>32.89</td>
<td>36.48</td>
</tr>
<tr>
<td>Separate $\ell_\infty$ + HRC</td>
<td>—</td>
<td>29.45</td>
<td>32.08</td>
</tr>
</tbody>
</table>

list of Figure 3. Following [6], we use 30 images per category for training, and others for testing.

We apply 3 versions of SPM models, i.e., a traditional BoF model (LLC [8]) for sparse coding, an improved model (GPP [6]) for spatial context modeling, and an ultimate version (EdgeGPP [6]) for combining heterogeneous features. Classification results are listed in Table 1.

4.2. The Caltech-UCSD Birds-200-2011 Dataset

The Caltech-UCSD Birds-200-2011 Dataset [19] contains 200 bird species and 11788 images in total. As shown in Figure 4, it is a very challenging fine-grained image collection. Following [10], we use 5 images per category to train the model, and test it on remaining images.

We apply 3 versions of HPM models learned from [10], with the difference of whether to learn a hierarchical part structure and whether to use Geometric Phrase Pooling. We name them LLC, LLC-HP (Hierarchical Part) and GPP-HP, respectively. Classification results are listed in Table 2.

4.3. Discussions

Here are some discussions based on experimental results.

- **Global $\ell_1$-norm vs. $\ell_2$-norm.** It is a popular opinion that $\ell_1$-normalization would cause the classification accuracies drop, especially in the models using max-pooling [8] [23]. However, we show in experiments that with a large enough coefficient $\nu$, global $\ell_1$-normalization gives comparable and even higher accuracies than global $\ell_2$ (see Table 2). i.e., hierarchical weighting beyond separate normalization.

- **Separate normalization vs. model selection.** The effect of separate normalization is highly related to model selection. SPM uses a naive spatial division ($4 \times 4$ grids), while HPM works on semantic body parts (beak, head, wings, tail, etc.). Comparatively speaking, HPM is more likely to preserve part-based properties, especially the independent assumption. That is why separate normalization produces much higher accuracies in HPM, but works slightly worse in SPM.

- **Hierarchical weighting.** Hierarchical weighting works better than equal weighting in both scenarios. Here, we benefit from the hierarchical-contribution assumption, and exploit a straightforward estimation to increase the weighting on more robust spatial bins.

In general, the choice of normalization strategy has a great impact on classification accuracies. The more we exploit prior knowledge, the better normalization results we could obtain.

5. CONCLUSIONS

In this paper, we claim the importance of feature normalization in part-based Bag-of-Features models, by representing novel algorithms, extensive experiments and in-depth discussions. Based on proper normalization power and coefficient, our approach differs from previous ones in the careful consideration of part-based properties, i.e., the independent assumption and the hierarchical-contribution assumption. Experimental results provide strong evidences to support our statements: (1) $\ell_1$-normalization is also good for image classification, and (2) hierarchical weighting beyond separate normalization is supreme in part-based BoF models.

6. ACKNOWLEDGEMENT

This work was supported by the National Basic Research Program (973 Program) of China under Grant 2012CB316301, and Basic Research Foundation of Tsinghua National Laboratory for Information Science and Technology (TNList). This work was also supported in part to Dr. Qi Tian by ARO grant W911NF-12-1-0057, NSF IIS 1052851, Faculty Research Awards by Google, NEC Laboratories of America and FXPAL, UTSA START-R award and NSFC 61128007.
7. REFERENCES


