Decomposing Bag of Words Histograms

Ankit Gandhi[1] Karteek Alahari[2,*] C. V. Jawahar[1]

[1]CVIT, IIIT Hyderabad, India [2]Inria, France

Abstract

We aim to decompose a global histogram representation of an image into histograms of its associated objects and regions. This task is formulated as an optimization problem, given a set of linear classifiers, which can effectively discriminate the object categories present in the image. Our decomposition bypasses harder problems associated with accurately localizing and segmenting objects. We evaluate our method on a wide variety of composite histograms, and also compare it with MRF*-based solutions. In addition to merely measuring the accuracy of decomposition, we also show the utility of the estimated object and background histograms for the task of image classification on the* PASCAL VOC 2007 *dataset.*

1. Introduction

There has been significant success in addressing the visual categorization problem in recent years. This can be attributed to advancements in feature descriptors (e.g. SIFT [17], HOG [5]), representations (e.g. Bag of Words (BOW) [29]), and classifiers (e.g. fast, scalable support vector machines (SVMs) [18, 33]). Often, success is measured in terms of increase in quantitative performance achieved on popular datasets such as Caltech [11] and PASCAL VOC [8]. In this context, the pipeline of BOW representation computed from dense SIFT (DSIFT) descriptors, followed by an SVM classifier has emerged as one of the most successful, as well as popular solutions for a wide spectrum of object and scene categories such as automobiles, rigid man-made objects, natural scenes [4, 32].

An SVM classifier is often trained to recognize only a single class category. When multiple objects (or uncorrelated noise) are present in an image, the performance deteriorates. To better understand this issue let us consider a split of the PASCAL VOC 2007 test data into images containing a single class category (PASCAL-S) and multiple class categories (PASCAL-M). In this setting, the average precision (AP) of the BOW-trained SVM classifier for the category "cat" is 0.589 on PASCAL-S, while only 0.189 on PASCAL-M. Also, it has been observed that BOW histograms of single isolated objects are relatively easy to classify. For example, accuracy as high as 77.78% is reported on Caltech 101 dataset [4], while more complex images, which contain multiple objects and natural clutter are harder to work with (e.g. 62.8% is still the best score on PASCAL VOC 2007 [10]). An important reason for this deterioration in performance is the fact that a classifier trained on single objects often fails to recognize the object when the global image representation (BOW) is "corrupted" by additional objects and clutter present in the image. A question of interest to us now is the following. *Is it possible to filter out the clutter and classify only the signal?*

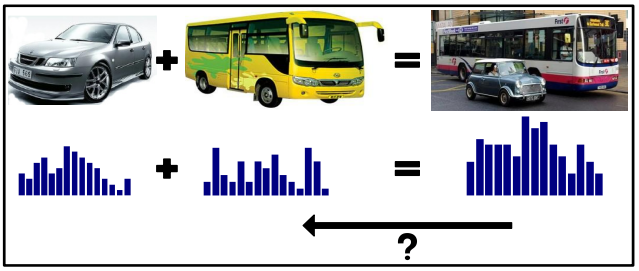

Figure 1: **Decomposing Bag of Words Histograms.** We are interested in obtaining the constituent histograms from a composite histogram.

In this work, our aim is to decompose a global BOW histogram into multiple histograms corresponding to different categories present in the image, as shown in Figure 1. We assume some prior knowledge about the possible categories present in the image, a superset of the class categories, for instance, and that each of these categories is learnt with an appropriate SVM classifier. We solve the problem by partitioning the image into regular cells and assigning weights that correspond to each of the categories. The weights are computed using a linear optimization scheme, while main-

*WILLOW project-team, Département d'Informatique de l'École Normale Supérieure, ENS/Inria/CNRS UMR 8548, Paris, France.

taining their spatial continuity. Thus, the histogram of each of the categories in the image can be computed using a weighted sum of the cell histograms. Our method is specially designed for feature representations, which are additive for an image, i.e. the feature representation of the image can be computed as the sum of the features corresponding to its associated regions.

Histogram decomposition has many applications, and can be used in multiple settings to boost the classification performance as we show in the experiments section — both when single and multiple categories are present in an image. The decomposition can also be used for separating object and background histograms in an image. We evaluate our method on images from various sources: Caltech-256, Flickr and PASCAL VOC 2007.

Related work. We note that existing approaches for object detection and semantic segmentation can be adapted to solve the histogram decomposition problem. This involves two steps: (i) Performing object detection or segmentation; and (ii) Computing the individual histograms for the classes using the bounding boxes or the segmentation masks obtained.

In the case of object detection, a classifier is run over an image at multiple scales and locations. Naturally, this process is computationally expensive. Many approaches have been proposed to overcome this by restricting the number of potential windows [15], segmenting the image [21], searching only the salient regions [1], sharing features across categories [31], speeding up the individual classifiers [30]. However, the more successful methods [9] still depend on an exhaustive search in the image space. On the other hand, segmenting an image, unsupervised [3, 26] as well as category-based [14, 27, 35], is a more complex problem, where the task is to obtain a (super)pixel-level labelling. In essence, using detection or segmentation approaches for solving the histogram decomposition problem would be an overkill. In this paper, we present a simpler and computationally efficient alternative for this problem.

The work of [28] uses pLSA statistical modelling to discover objects. They represent an image as a mixture of topics, and compute a histogram from a mixture of histograms corresponding to each topic. Verbeek *et al*. [34] build on this by introducing a spatial coherency of labels for region classification. The recent works of [23, 25] are more closely related to our work. Russakovsky *et al*. [23] use the intuition that performing localization and classification simultaneously can be beneficial. They first infer the location of an object of interest and then pool low-level features separately in the foreground and background regions to form an image-level representation. This can be viewed as decomposing the image into object and background using an object detection method. Although this is an interesting approach, it suffers from high computational costs both for training and testing. Nevertheless, we compare the decomposition obtained by their work with ours in Section 4. It must be noted that our goal is not to achieve an exact localization of objects, unlike that in [23]. The work of Sharma *et al*. [25] on spatial saliency also partitions the image into regular cells and assigns weights to them. As claimed in their paper, it is more suitable for fine-grained and scene classification, and less so for classification of objects – a task we consider in this work. Furthermore, it does not consider the spatial continuity of weights while assigning them to cells as we do. The significance of this continuity term for classification of objects is further discussed in Section 4.

The remainder of the paper is organized as follows. We formulate our task as an optimization problem in Section 2. We also contrast our approach with MRF-based methods, and discuss the latter's limitation in Section 2. Inspired by the success of fast and scalable classifiers [18, 33, 37], we discuss our work using linear SVM as an example. We show how the formulation can be generalized to spatially-constrained decomposition in Section 3. Section 4 presents an exhaustive evaluation of our method. We then make concluding remarks in Section 5.

2. Decomposing Histograms

Consider an image with k object classes of interest. Let $\mathbf{h}$ denote the global unnormalized[1] histogram of the image. The set of linear classifiers (e.g. SVM) trained on the k classes are represented by $\mathbf{w}_1, \ldots, \mathbf{w}_k$. Our objective is then to decompose the histogram $\mathbf{h}$ into k constituent histograms represented by $\mathbf{x}_1, \ldots, \mathbf{x}_k$, corresponding to each of the classes.

To solve the histogram decomposition problem, we begin by partitioning the image into $M \times N$ regular rectangular cells. Let $\mathbf{h}_{ij}$ denote the histogram computed independently for each cell. We introduce a binary variable $b_{ij}^p \in \{0, 1\}$ for each cell to denote whether it is part of an object from the pth category or not. With this, our decomposition problem is formulated as:

$$\max \sum_{p=1}^{k} \mathbf{w}_p^T \mathbf{x}_p, \tag{1}$$

such that $\mathbf{x}_p = \sum_{ij} b_{ij}^p \mathbf{h}_{ij}$ and $\sum_p b_{ij}^p = 1$. Note that $\mathbf{w}_p^T \mathbf{h}_{ij}$ can be compared for different classes since the classifiers ($\mathbf{w}_p$'s) are trained on normalized histograms. This problem can be solved in closed form by taking b_{ij}^p to be 1 for the p that maximizes $\mathbf{w}_p^T b_{ij}^p$ and 0 for all other p's.

The optimal solution to the problem (1), however, is not always semantically meaningful. For instance, cells from

[1] Such histograms have been used successfully in the past [15].

sky or road may be labelled as part of other object categories such as bus or car. Furthermore, object cells of a specific category may be scattered and spatially disconnected. We also need a mechanism to incorporate some of the prior knowledge one may have in many practical situations. Examples of such constraints include: (i) a specific shape or aspect ratio of an object, (ii) spatial continuity and penalizing configurations that result in a set of scattered cells for an object category, (iii) a prior on the scale of the object of interest, or even (iv) favouring objects in certain parts of the image, say in the center. To make the formulation more realistic, we relax the assumption in (1) that all the cells are to be assigned to one of the k objects of interest. This implicitly accounts for the fact that not all parts of the image may be recognized with existing classifiers $\mathbf{w}_p$. Thus, the problem is to:

$$\max_{b,\lambda} \sum_{p=1}^{k} \mathbf{w}_p^T \mathbf{x}_p - \gamma \sum_{(i,j)\in\mathcal{N}} |\lambda_{ij}|, \tag{2}$$

subject to the constraints:

$$A : \mathbf{x}_p = \sum_{i,j} b_{ij}^p \mathbf{h}_{ij}, \qquad B : \sum_p b_{ij}^p \leq 1,$$

$$C : \sum_p \sum_{ij} b_{ij}^p \geq \mathbf{P}, \qquad D : b_{ij}^p \in \{0,1\},$$

$$E_1 : b_{ij}^p - b_{i+1,j}^p = \lambda_{i+1,j}, \quad E_2 : b_{ij}^p - b_{i,j+1}^p = \lambda_{i,j+1},$$

where γ is a regularization parameter. The constraint A defines an object class category as a weighted sum of histograms from multiple cells, similar to (1). The constraint B allows some of the cells to remain unlabelled. We introduce constraints E_1 and E_2 in a neighbourhood system $\mathcal{N}$, which enforce neighbouring cells to take a similar label.[2] In other words, E_1 and E_2 define penalties λ_{ij} and provide object smoothness constraints. The parameter γ controls the emphasis on the spatial smoothness of the object. The SVM classifiers used in this formulation do not include a bias term, but it can be easily incorporated by augmenting every histogram $\mathbf{x}_p$ with 1. Empirically, we found the effect of introducing bias negligible, as we are using unnormalized histograms $\mathbf{h}_{ij}$. The no-bias formulation favours discarding assignments with negative dot product ($\mathbf{w}_p^T \mathbf{x}_p$). However, the constraint C helps overcome this issue by enforcing a minimum of label assignments.

A familiar analogy of the objective function (2) with constraints A, B, D, E_1, E_2, are energy functions that can be modelled as a Markov random field (MRF) with unary and pairwise potentials, and solved efficiently [13]. In such a setting, the term $\mathbf{w}_p^T \mathbf{x}_p$ in (2) corresponds to the unary potential and the term $|\lambda_{ij}|$ represents the pairwise potential. However, as discussed in Section 2.1, the constraint C, which introduces a lower bound P on the number of cells assigned to any of the k classes, cannot be easily added into this framework. This constraint avoids a trivial solution for (2), and we will study the importance of this constraint in Section 4. Note that the residual histogram, i.e. $\mathbf{r} = \mathbf{h} - \sum_{i=1}^{k} \mathbf{x}_i$, need not be empty by design. If there is a need to use context for enhancing the representation, $\mathbf{r}$ can be added to a specific category. We relax the constraint D as $b_{ij}^p \in [0,1]$ and solve the resulting linear program (LP) relaxation. The spatial extents of the individual constituent histograms can be obtained by rounding b_{ij}^p's to their nearest integers.

The histograms of different categories in an image can be obtained directly using the solution of the LP relaxation, i.e. taking the weighted sum of the cell histograms (LP-relax) or by first rounding-off the solution to the nearest integer and then adding the corresponding cell histograms (LP-round). We analyze the performance on these two solutions in the experiments section, and observe that LP-relax performs better than LP-round, as the former makes a soft-assignment of cells to categories.

2.1. An MRF-based solution

As noted earlier, the decomposition problem can be modelled as an MRF energy minimization problem. Here, each cell in the $M \times N$ grid is represented as a node in a graph. Each node takes a label from the set $\mathcal{L} = \{1, \ldots, k\}$. This is equivalent to introducing binary variables b_{ij}^p for all the classes, and constraining them with B. Popular techniques such as sequential tree-reweighted message passing (TRW-S) [12], belief propagation [22], and alpha expansion [2] can be used to solve this formulation, which is equivalent to problem (2) with constraints A, B, D, E_1 and E_2. We observed (see Section 4) that most of the cells are assigned to the background in this solution. This is not surprising as the classifier used in the formulation is learnt from images with large intra-class variations, and is not expected to fire positively when only some part of the object is considered in a cell. We overcome this issue by introducing the constraint C. We study the effect of its introduction in Section 4 by solving the LP and MRF problems in the absence of C. Global constraints such as C cannot be incorporated easily into a standard pairwise MRF [6, 36]. Although MRFs with higher order potentials can be adapted to do so, they often lead to approximate and computationally expensive solutions. The work of [24] introduces a count based global prior constraint, similar to C, into an MRF formulation, but resorts to LP-based techniques to solve the problem. Further, it focusses on obtaining integral solutions for segmenting an image, whereas the solution of the LP relaxation suffices for our histogram decomposition task.

[2] Note that constraints E_1 and E_2 exist for every neighbouring pair of cells. A subset of constraints is shown here as an example. We consider constraints of neighbourhood size 8 in our formulation.

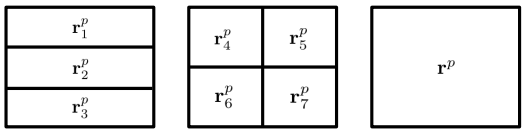

Figure 2: Incorporation of spatial pyramid histograms into our LP formulation. We show the sub-regions considered while computing the spatial histogram for an object category p. In addition to the constraints mentioned here, we add constraints that $\mathbf{r}_1^p$, $\mathbf{r}_2^p$ and $\mathbf{r}_3^p$ should contain equal number of visual words, as they occupy the same area in the image, and a similar constraint for $\mathbf{r}_4^p, \mathbf{r}_5^p, \mathbf{r}_6^p$ and $\mathbf{r}_7^p$.

3. Spatially-constrained Decomposition

Our formulation discussed thus far encodes limited spatial information of the objects or the regions into the histograms. Often, incorporating spatial information, such as in [16], and concatenating histograms from multiple sub-regions, has shown to improve the classification performance in many cases. In this section we present a more general framework to address this issue.

We extend our framework introduced in (2) to incorporate spatial information into histograms. We introduce weak geometry constraints into the histograms, without affecting the linearity of the problem, inspired by the work on spatial histograms [16]. The spatial region for an object category *p* is divided into 3×1, 2×2 and 1×1 grids giving rise to a total of eight sub-regions, as shown in Figure 2, similar to [4]. The final representation of an object is obtained by concatenating histograms of the eight sub-regions. Let $\mathbf{r}^p$ be the histogram of class p and $\mathbf{r}_1^p, \ldots, \mathbf{r}_7^p$ denote the corresponding sub-region histograms. In this formulation involving Spatial Pyramid Matching (SPM), we simultaneously solve for b_{ij}^p and sub-region histograms $\mathbf{r}_1^p, \ldots, \mathbf{r}_7^p$. We replace the constraint A in problem (2) with the following set of constraints:

$$A1 : \mathbf{x}_p = [\mathbf{r}_1^p \ldots \mathbf{r}_4^p \ldots \mathbf{r}_7^p \; \mathbf{r}^p], \quad A2 : \sum_{ij} b_{ij}^p \mathbf{h}_{ij} = \mathbf{r}^p,$$

$$A3 : \sum_{i=1}^{D} \mathbf{r}_{1i}^p = \ldots = \sum_{i=1}^{D} \mathbf{r}_{3i}^p, \quad A4 : \sum_{k=1}^{3} \mathbf{r}_k^p = \mathbf{r}^p,$$

$$A5 : \sum_{i=1}^{D} \mathbf{r}_{4i}^p = \ldots = \sum_{i=1}^{D} \mathbf{r}_{7i}^p, \quad A6 : \sum_{k=4}^{7} \mathbf{r}_k^p = \mathbf{r}^p,$$

where D is the dimension of the histogram, $\mathbf{r}_j^p$ is the histogram of the jth sub-region, and $\mathbf{r}_{ji}^p$ denotes the ith visual word count in the histogram $\mathbf{r}_j^p$. Constraint $A1$ defines the histogram of object p as the concatenation of eight sub-region histograms shown in Figure 2. The constraint $A2$ represents the histogram in terms of its cell histograms. Constraints $A3$ and $A5$ imply that $\mathbf{r}_1^p$, $\mathbf{r}_2^p$, $\mathbf{r}_3^p$ and $\mathbf{r}_4^p$, $\mathbf{r}_5^p$ $\mathbf{r}_6^p$, $\mathbf{r}_7^p$ contain equal number of visual words, as they occupy the same area in the image. Constraints $A4$ and $A6$ correspond to the conditions mentioned in Figure 2, that the sum of the sub-region histograms is equal to the histogram of class p. Note that only weak spatial constraints for sub-region histograms have been considered in the above formulation, so as to keep our problem linear. It does not impose constraints on the positions of sub-regions in the image, and hence, does not compute exact spatial histograms. However, it encodes some spatial information, which results in a better decomposition of the global histogram.

We use SVMs trained on spatial histograms of tight bounding boxes around the object as classifiers ($\mathbf{w}_p$), computed by dividing object bounding box into 3×1, 2×2 and 1×1 grids, as shown in the illustration in Figure 2. Hence, when we maximize $\mathbf{w}_p^T \mathbf{x}_p$ in the objective function, $\mathbf{r}_1^p$, $\ldots, \mathbf{r}_7^p$ approximately correspond to the histograms of sub-regions assumed.

4. Experiments and Results

We demonstrate the performance of our method for decomposing the global BOW histogram of an image into its constituent histograms in a variety of settings. In addition to merely measuring the accuracy of the decomposition, we also show its utility for the task of image classification. We use PASCAL VOC 2007, Flickr multiple object, and Caltech-256 based datasets in our experiment.

Composed Caltech dataset. The Composed Caltech dataset, we refer to as CALTECH, is a synthetic dataset generated from Caltech-256 images [11]. It consists of images formed by "pasting" scaled, translated, and rotated versions of the original Caltech-256 images onto other images from the dataset. This provides us with a controlled setting to regulate the complexity of the image histograms. Note that the visual content of some of these images may not be appealing, but we are only interested in the histograms. We divide the images from CALTECH into three multi-category object datasets – CALTECH-2, CALTECH-3 and CALTECH-4 – each consisting of 10,000 images. They have been created by concatenating two, three, and four objects, randomly selected from 20 predefined candidate categories, respectively. The scale of each object in the composite image is measured as the percentage of the composite histogram it contributes to. This varies from 10% to 90%. The purpose of introducing this dataset was to study the sensitivity and robustness of our formulation especially when the object size in the image, and k, the number of categories considered in the objective function (2), vary.

PASCAL VOC and Flickr multiple object dataset. We also use natural images from PASCAL VOC 2007 [7] and a

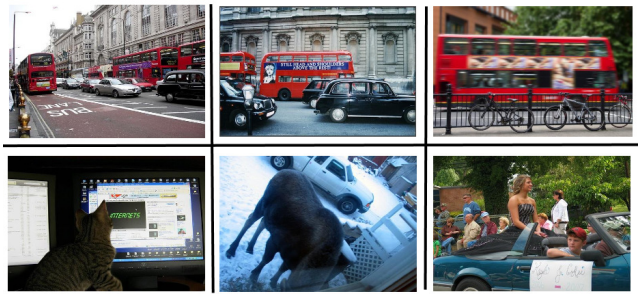

Figure 3: Example images from Flickr (first row) and PASCAL VOC 2007 (second row), containing multiple objects.

set of images downloaded from Flickr. The Flickr dataset is composed of images downloaded with "*bus* & *car*" and "*bus* & *bicycle*" text queries, which were then filtered manually. We refer to the Flickr multiple object datasets as Flickr-M1 and Flickr-M2. Flickr-M1 has 196 positive images containing both *bus* & *car*, and Flickr-M2 has 209 positive images with both *bus* & *bicycle* in them. For both these datasets, we harvest negative training examples from the PASCAL VOC 2007 images containing neither of the two object categories. A few samples from these datasets are shown in Figure 3.

Experimental setting. In all our experiments an image is divided into 16×16 cells, and a vocabulary of size 4K is used, unless otherwise stated. DSIFT features are extracted at a step size of 5 pixels and a hard quantization is used to assign them to visual words. We trained SVM classifiers using liblinear. For the CALTECH dataset experiments, the classifiers are trained using 25 samples from each category that are not in the CALTECH dataset. We set P (from constraint C in (2)) to 50% of the total cells in an image. The parameter γ is set to 1 for CALTECH and 0.7 for Flickr and PASCAL VOC 2007 datasets by cross validation. We used MOSEK[3] for solving the linear programs.

4.1. CALTECH histogram decomposition

Comparison of LP and TRW-S. We begin by evaluating the performance of our histogram decomposition method on the CALTECH dataset. Recall that the problem (2), albeit without the inclusion of constraint C, can be solved: (i) directly as an LP formulation; or (ii) using an MRF-based solution (Section 2.1) such as sequential tree-reweighted message passing (TRW-S) [12]. We follow both these approaches on BOW histograms of the CALTECH dataset. Since we use DSIFT-BOW histograms, the orientation and location of the individual objects do not significantly affect the composite histogram. We evaluate the performance of our decomposition by obtaining the mean AP over all the classes, when the constituent (category-level) histograms are passed to the respective SVM classifiers. Table 1 shows

[3]http://www.mosek.com

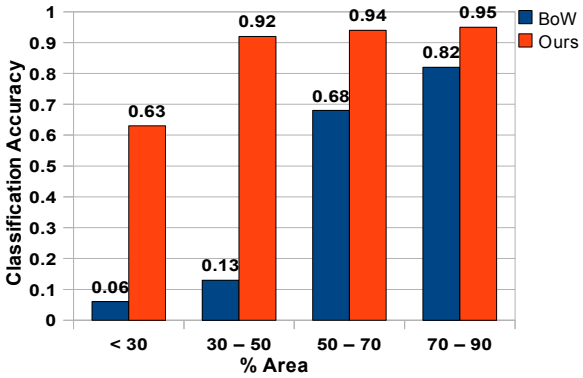

Figure 4: Comparison of the classification performance and the scale of objects in the CALTECH dataset. We use *k*=5 for this experiment. We observe that our method outperforms the naive BOW based classifiers at all scales.

the MAP obtained using LP (LP (w/o C)) and TRW-S, without constraint C. It can be observed that the performance of these two solution schemes is comparable, with LP (w/o C) performing better in some cases, and marginally inferior in some other cases.

Importance of the constraint C. The advantage of using an LP-based solution is evident with the inclusion of the constraint C. Table 1 also shows the MAP over all the 20 Caltech classes when the problem is solved using the LP formulation with the constraint C, i.e. (2), referred to as LP (with C) in the table. It shows a significant boost in performance over TRW-S and LP (w/o C) – an average improvement of 30.35%. Note that including this global constraint in the TRW-S solver is not trivial, as discussed in Section 2.1. We also compare the LP method with two other baseline approaches: BOW and cell-based voting (CV). We use the entire composite histogram of an image for the BOW method, while for the CV approach we assign each cell independently to atmost one object, and build the object histogram from histograms of cells that belong to the object. Our decomposition based methods outperform the baselines, as shown in Table 1. For the remainder of the experiments, we used the LP formulation with the constraint C.

Figure 4 analyzes the classification accuracy when the scale of the object (percentage of the area occupied) in the image varies. As expected, the problem is relatively hard when the object is very small. When the scale is 30% or more, decomposition helps classify the image more than 92.4% of the time. Even when the scale of the object is small (10-30%), our method correctly discriminates the object histograms more than 63% of the time. Note that the improvement of our method over BOW is more pronounced when the scale is small. This is significant since, it is these small objects that we often find hard to recognize in composite images, in the presence of clutter.

k	CALTECH-2			CALTECH-3			CALTECH-4		
	LP (w/o C)	TRW-S	LP (with C)	LP (w/o C)	TRW-S	LP (with C)	LP (w/o C)	TRW-S	LP (with C)
BoW	0.19			0.23			0.22		
CV	0.23			0.25			0.26		
2	1.00	1.00	**1.00**	NA	NA	NA	NA	NA	NA
3	0.78	0.80	**0.97**	0.92	0.93	**0.98**	NA	NA	NA
4	0.61	0.63	**0.89**	0.79	0.80	**0.94**	0.84	0.86	**0.97**
5	0.51	0.54	**0.76**	0.69	0.72	**0.81**	0.82	0.82	**0.92**
10	0.29	0.25	**0.41**	0.35	0.37	**0.45**	0.36	0.37	**0.39**
20	0.15	0.16	**0.26**	0.17	0.17	**0.29**	0.19	0.21	**0.31**

Table 1: **Comparison of LP (w/o C), TRW-S & LP (with C).** Mean classification AP for different k's on CALTECH using TRW-S and our LP-based solution (with and without the constraint C). CV and BoW are two baseline methods. BoW uses the entire composite image histogram and CV uses histograms obtained via cell-based voting. Note that the formulation is not solved (NA) for 2 classes ($k = 2$) when 3 or 4 objects are present in the image.

Figure 5: Histogram decomposition on Flickr-M2 dataset. LP is solved for two classes *bus* and *bicycle* simultaneously. The images and the corresponding weights obtained for their cells using the LP solution are shown. The cells shown in red are weights of *bus*, while those in green are of *bicycle*. Higher intensity of the colour represents a value closer to 1. (**Best viewed in colour.**)

4.2. Multiple object classification

In this experiment we investigate how the presence of one object in an image can negatively affect the classification of others. We consider the AP obtained when the entire histogram of an image is given to an SVM classifier (BoW) as the baseline. Using the LP formulation proposed in Section 2, we split the image histogram into histograms of constituent objects, and the background/context. Figure 5 shows the decomposition of the global histograms in a few examples containing bus and $bicycle$ categories. One approach to evaluate this decomposition is by using the constituent histograms directly in an object classifier. However, ignoring the background/context histogram would be an unwise move, given previous work [19] suggesting that background/context can provide strong cues useful for classifying objects. Thus, we use the object-background feature representation of [23], where a histogram is represented by a concatenation of object and background histograms.

Method	Flickr-M1		Flickr-M2		PASCAL-M
	bus	*car*	*bus*	*bic*	
BoW	0.522	0.516	0.302	0.108	0.287
LP-round	0.561	0.574	0.373	0.235	0.312
LP-relax	0.582	0.590	0.397	0.246	0.331
LP-SPM	**0.598**	**0.612**	**0.408**	**0.278**	**0.348**

Table 2: Classification AP on Flickr-M1, Flickr-M2, and PASCAL-M datasets. In LP-relax, we use soft assignment of cells whereas in LP-round, hard assignment of cells is used.

The background histogram is obtained by subtracting histograms of objects from the global image histogram.

We compute AP on the Flickr dataset with classifiers trained on features extracted from object bounding boxes, concatenated with the features extracted from the remainder of the image. The classifiers used in our LP formulation are trained on features extracted from the training+validation sets of PASCAL VOC 2007. LP is solved for all images in the dataset to get the constituent histograms (for Flickr-M1, it is solved using classifiers for *bus* and *car*, and for Flickr-M2, using classifiers for *bus* and *bicycle*). Table 2 compares the classification AP obtained using BoW, LP and LP-SPM (Section 3). Using LP, we see an improvement of 12.9% on Flickr-M1, and 56.8% on Flickr-M2. The relaxed LP solution performs better than the rounded-off solution, as it does not make hard assignments for cells. Table 2 also compares the AP on PASCAL-M, with LP showing the best results.

4.3. Decomposition into object and background

We now discuss the decomposition results on the PASCAL VOC 2007 dataset. Figure 6 shows the assignment of object *vs*. background labels to the image cells on a few sample images from the dataset. This decomposition is evaluated in the context of the image classification problem.

Table 3 shows a comparison of our LP decomposition scheme with baseline methods. The image representation

Figure 6: An illustration of the decomposition results on sample images from PASCAL VOC 2007 dataset. We show an image and the corresponding weights of its cells obtained from our LP solution. Higher intensity of the colour green represents a value closer to 1. (**Best viewed in colour.**)

Method	PASCAL VOC		Comments
	Simple	SPM+EX.F.M.	
BOW	0.379	0.528	Baseline
TestBB	0.659	0.834	Golden baseline
DPM	0.386	0.543	Decomp. using existing methods
Sem. Seg.	0.434	0.561	
LP-round	0.418	0.536	Decomp. using LP-based methods
LP-relax	0.435	0.558	
LP-SPM	**0.447**	**0.567**	

Table 3: Mean classification AP on PASCAL VOC 2007 over all the 20 classes. AP for each of the 20 classes can be found on the project website. TestBB shows the AP when the decomposition is done using ground truth bounding boxes, DPM when using [9], Sem. Seg. is with ALE [14], and LP is the proposed formulation. See text for details.

is a concatenation of individually normalized object and background histograms for the methods TestBB, DPM, Sem. Seg. and LP, while a normalized histogram of the entire image is used in the case of BOW. The representation is then used by the SVM classifier. The image classification AP is computed for all the methods mentioned in Table 3, with (SPM + EX.F.M.) and without (*Simple*) the use of spatial pyramids and explicit feature maps [4, 33]. Spatial features are computed as in [4], both for the object and the background regions. In the table, BOW refers to a direct baseline when the entire image histogram is used. TestBB is the "golden" baseline, where the object histograms are extracted from ground truth bounding boxes, and used in combination with histograms from the remainder of the image. This provides us with an upper bound on this dataset for an object-background representation model. We also compared our approach with methods using regions from sliding window detectors [9] (DPM) and semantic segmentation [14] (Sem. Seg.). For DPM, we use the publicly available detectors to get the bounding boxes, and for semantic segmentation, we use the automated labelling environment (ALE) framework to find the segmentation mask [14].[4] The features from object regions are concatenated with features from the remainder of the image, as done for other methods, for both these approaches. We observe that our LP decomposition scheme outperforms DPM and and is comparable to Sem. Seg., with a much lower computation cost (less than 0.1s). Although the running time of DPM and Sem. Seg. approach can be improved significantly by applying them at a coarse resolution, this results in an inferior AP. For example, down sampling PASCAL VOC 2007 images by a fourth reduces the performance of (re-trained) ALE from 0.567 in Table 3 to 0.288.

4.4. Decomposition in a weakly supervised setting

In the histogram decomposition formulation (2), we require a linear SVM classifier, which can discriminate the categories present in the image. We learn them using bounding boxes annotated in the training data. We now extend our approach to learn classifiers for a more general weakly supervised setting, where bounding box annotation is not available for most of the images. Recent approaches, such as [20, 23], can be adapted to learn classifiers in a weakly supervised setting, but being based on object localization, they are computationally expensive.

We use an iterative procedure to learn the classifier in this setting. Given an initial classifier for an object, we decompose histograms of training images with our LP formulation. Next, we compute the new histograms of objects as a weighted sum of the cell histograms and re-train the classifiers with them. We repeat this procedure 10 times. In order to initialize, we use 10 ground truth bounding box annotations per object and train an initial classifier. We compare this decomposition scheme for the image classification task on PASCAL VOC 2007 to object-centric spatial pooling [23]. We use DSIFT at a step size of 3 and sum pooling based LLC encoding (to ensure additivity of features) while

[4] We trained ALE on ($\sim$ 2223) images from PASCAL VOC 2011 and not on the ($\sim$ 422) images from PASCAL VOC 2007.

solving LP. Following [23], we use max pooling based LLC encoding for the object-background feature representation when classifying images. We achieve a mean AP of 0.560 with LP-relax and 0.571 with LP-SPM, using a vocabulary of size 4k compared to 0.572 in [23]. In the same setting, when we increase the vocabulary size to 25k, we obtain an AP of 0.589 for LP and 0.594 for LP-SPM. The training time on a single machine for the method in [23] is 3 days, compared to under 7 hours for our LP method. It must be noted that the AP scores mentioned in Table 3 are not directly comparable to those given in this section, as the method for pooling is based on [23], and different to those used previously. The features used and the vocabulary size are different as well. We also analyzed the importance of the spatial smoothness constraint by setting $\gamma = 0$, which results in an AP of 0.458 compared to 0.560 for LP-relax.

5. Summary

We proposed an effective method to decompose a global histogram of an image into histograms of its associated objects and regions. Our approach solves the problem using an LP formulation, by taking an intermediate path between two harder problems, namely bounding box accurate object detection and pixel-accurate object segmentation. We showed that a wide variety of composite histograms can be decomposed into their constituent histograms with our LP method. We also demonstrated the application of histogram decomposition for improving the classification performance on multiple object and PASCAL VOC 2007 datasets using an object-background representation of an image.

Acknowledgements. The authors would like to thank Andrew Zisserman for helpful suggestions. Karteek Alahari is partly supported by the Quaero programme funded by the OSEO.

References

[1] B. Alexe, T. Deselaers, and V. Ferrari. What is an object? In *CVPR*, 2010.

[2] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *PAMI*, 2001.

[3] Y. Chai, V. Lempitsky, and A. Zisserman. BiCoS: A bi-level co-segmentation method for image classification. In *ICCV*, 2011.

[4] K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman. The devil is in the details: an evaluation of recent feature encoding methods. In *BMVC*, 2011.

[5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[6] A. Delong, A. Osokin, H. N. Isack, and Y. Boykov. Fast Approximate Energy Minimization with Label Costs. In *CVPR*, 2010.

[7] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results.

[8] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (VOC) challenge. *IJCV*, 2010.

[9] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *PAMI*, 2010.

[10] B. Fernando, E. Fromont, and T. Tuytelaars. Effective use of frequent itemset mining for image classification. In *ECCV*, 2012.

[11] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. Technical Report 7694, California Institute of Technology, 2007.

[12] V. Kolmogorov. Convergent tree-reweighted message passing for energy minimization. *PAMI*, 2006.

[13] V. Kolmogorov and R. Zabih. What energy functions can be minimized via graph cuts. *PAMI*, 2004.

[14] L. Ladicky, C. Russell, P. Kohli, and P. H. S. Torr. Associative hierarchical crfs for object class image segmentation. In *ICCV*, 2009.

[15] C. H. Lampert, M. B. Blaschko, and T. Hofmann. Beyond sliding windows: Object localization by efficient subwindow search. In *CVPR*, 2008.

[16] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.

[17] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2004.

[18] S. Maji, A. C. Berg, and J. Malik. Classification using intersection kernel support vector machines is efficient. In *CVPR*, 2008.

[19] A. Oliva and A. Torralba. The role of context in object recognition. *Trends in Cognitive Sciences*, 2007.

[20] M. Pandey and S. Lazebnik. Scene recognition and weakly supervised object localization with deformable part-based models. In *ICCV*, 2011.

[21] O. M. Parkhi, A. Vedaldi, C. V. Jawahar, and A. Zisserman. The truth about cats and dogs. In *ICCV*, 2011.

[22] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.

[23] O. Russakovsky, Y. Lin, K. Yu, and L. Fei-Fei. Object-centric spatial pooling for image classification. In *ECCV*, 2012.

[24] T. Schoenemann. Minimizing count-based high order terms in markov random fields. In *EMMCVPR*, 2011.

[25] G. Sharma, F. Jurie, and C. Schmid. Discriminative spatial saliency for image classification. In *CVPR*, 2012.

[26] J. Shi and J. Malik. Normalized cuts and image segmentation. *PAMI*, 1997.

[27] D. Singaraju and R. Vidal. Using global bag of features models in random fields for joint categorization and segmentation of objects. In *CVPR*, 2011.

[28] J. Sivic, B. Russell, A. Efros, A. Zisserman, and W. Freeman. Discovering objects and their location in images. In *ICCV*, 2005.

[29] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *ICCV*, 2003.

[30] H. O. Song, S. Zickler, T. Althoff, R. Girshick, M. Fritz, C. Geyer, P. Felzenszwalb, and T. Darrell. Sparselet models for efficient multiclass object detection. In *ECCV*, 2012.

[31] A. Torralba, K. P. Murphy, and W. T. Freeman. Sharing visual features for multiclass and multiview object detection. *PAMI*, 2007.

[32] A. Vedaldi and B. Fulkerson. VLFeat: An open and portable library of computer vision algorithms, 2008.

[33] A. Vedaldi and A. Zisserman. Efficient additive kernels via explicit feature maps. *PAMI*, 2011.

[34] J. Verbeek and B. Triggs. Region classification with markov field aspect models. In *CVPR*, 2007.

[35] J. Winn and J. Shotton. The layout consistent random field for recognizing and segmenting partially occluded objects. In *CVPR*, 2006.

[36] O. J. Woodford, C. Rother, and V. Kolmogorov. A global perspective on map inference for low-level vision. In *ICCV*, 2009.

[37] J. Wu. Power mean SVM for large scale visual classification. In *CVPR*, 2012.