This CVPR2014 paper is the Open Access version, provided by the Computer Vision Foundation.
The authoritative version of this paper is available in IEEE Xplore.

NMF-KNN: Image Annotation using Weighted Multi-view Non-negative Matrix Factorization

Mahdi M. Kalayeh Haroon Idrees Mubarak Shah
Center for Research in Computer Vision, University of Central Florida
{mahdi, haroon, shah}@eecs.ucf.edu

Abstract

The real world image databases such as Flickr are characterized by continuous addition of new images. The recent approaches for image annotation, i.e. the problem of assigning tags to images, have two major drawbacks. First, either models are learned using the entire training data, or to handle the issue of dataset imbalance, tag-specific discriminative models are trained. Such models become obsolete and require relearning when new images and tags are added to database. Second, the task of feature-fusion is typically dealt using ad-hoc approaches. In this paper, we present a weighted extension of Multi-view Non-negative Matrix Factorization (NMF) to address the aforementioned drawbacks. The key idea is to learn query-specific generative model on the features of nearest-neighbors and tags using the proposed NMF-KNN approach which imposes consensus constraint on the coefficient matrices across different features. This results in coefficient vectors across features to be consistent and, thus, naturally solves the problem of feature fusion, while the weight matrices introduced in the proposed formulation alleviate the issue of dataset imbalance. Furthermore, our approach, being query-specific, is unaffected by addition of images and tags in a database. We tested our method on two datasets used for evaluation of image annotation and obtained competitive results.

1. Introduction

Image annotation refers to the task of assigning relevant tags to query images based on their visual content [22, 15]. The problem is difficult because an arbitrary image can capture a variety of visual concepts, each of which would require separate detection. Each image can be represented using multiple features which may be low-level, e.g. RGB histograms and HOG, or mid-level such as object concepts, e.g. human, dog, sky etc., or even high-level denoting the broader class to which the image belongs, e.g., structures, animal, food. These different features capture different aspects or views[1] of the image, thereby, providing complementary information. However, since each feature represents the same image, they all capture the same underlying latent structure. That is, it is possible to transform feature vectors for each image so that the new representations, with respect to some pre-defined distance metric, are consistent across all the views.

Automatic image annotation is crucial for searchable databases like Flickr, Photobucket, Picassa or Facebook. One of the key characteristics of real world databases is the continuous addition of new images, which contain new tags as well. Till 2011, 6 billion images had been uploaded to Flickr[2] while almost a quarter trillion images have been shared on Facebook[3] with a total of 300 million images uploaded every day. For a method to be practical for such databases, it has to rely on minimal training as the addition of new images and tags can render the learned models less effective over time. This holds true for both the methods that learn a direct mapping from features to tags [38, 3], or those that learn tag-specific discriminative models [15, 30, 34] where positive set contains images which contain a particular tag and the negative set contains images which do not have that tag. Obviously, as new images and tags are introduced into the database, the positive set for each tag will change, requiring retraining of the models.

Inspired by the success of the recent nearest-neighbor approaches for image annotation [22, 15], we propose a novel method that learns a query-specific generative model using the nearest-neighbors. The proposed approach is illustrated in Figure 1. The key idea of our approach is to treat tags as another view in addition to visual features, and find a joint factorization of all views into basis and coefficient matrices such that the coefficients of each training image are similar across views. This forces each basis vector to capture same latent concept in each view as well. After the factorization, the tags are transferred using both the model (basis) and the

[1] For consistency with Machine Learning literature, views means features in this paper.
[2] http://tinyurl.com/q4zdshq
[3] http://tinyurl.com/pktnba2

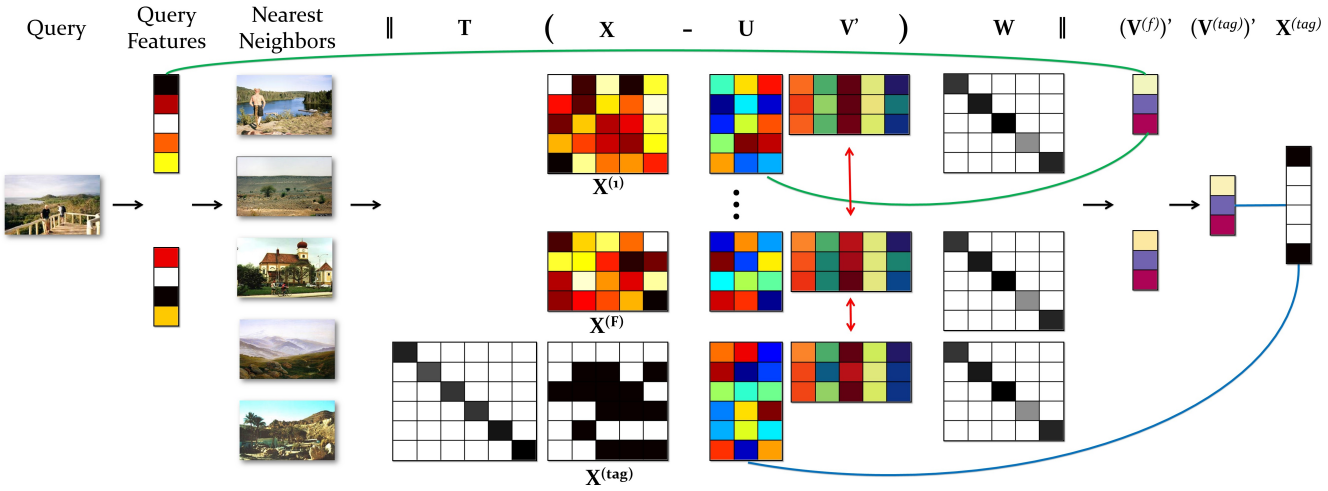

Figure 1. Schematic illustration of the proposed method: Given a query image, we extract different features which are used to find its nearest-neighbors. Then, using Non-negative Matrix Factorization using all the features $\mathbf{X}$, including tags, we find basis $\mathbf{U}$ and coefficient matrices $\mathbf{V}'$. The factorization for all matrices is done in a joint fashion by imposing a consensus constraint (red double arrows). Furthermore, to handle dataset imbalance, we introduce weight matrices $\mathbf{T}$ and $\mathbf{W}$ within the formulation. Using the basis matrices and corresponding features of the query, we find coefficient vector for each view (green lines). The matrix product of the tag-basis $\mathbf{U}^{(tag)}$ and mean of coefficient vectors from all views gives score for each tag (blue lines).

visual features of the query. Thus, given a query image, we first extract different visual features and find its nearest-neighbors. Then, using Non-negative Matrix Factorization on all views, i.e., visual features as well as tags, we find basis and coefficient matrices. The coefficient vectors from all views, recovered using visual features of the query and the corresponding basis, are then averaged to get a unique coefficient vector. The matrix product of the final vector with basis for tags gives the scores for individual tags.

Non-negative Matrix Factorization [19] is a well-studied problem where the aim is to decompose a matrix into non-negative basis and coefficient matrices. The non-negative coefficients can then be seen as a soft assignment in terms of discovered basis [10, 5]. For the case of image annotation where multiple views are available, this problem requires NMF across all views. This is severely under-constrained and since the views capture the same latent structure, a consensus regularization can enforce the solution to discover a consistent latent structure for all views [21]. Each basis in each view then represents the same latent concept across views. In this work, we treat annotated tags as another view of the image and learn a set of basis across all views that correspond to the same underlying concepts. Note that, these concepts may not have any semantics associated with them, all that is implied is consistency in terms of the abstraction they capture.

One important issue in image annotation as encountered by all previous works is that of rare tags - tags that do not occur frequently in the training data. For that, we introduce two weight matrices within the Multi-view NMF framework that increase the importance of both the rare tags and the images that contain rare tags. By assigning suitable weights, the NMF learns consistent latent concepts that are forced to capture the rare tags well, thus, allowing us to alleviate the issue of dataset imbalance by increasing recall for the rare tags. In summary, we propose to use Multi-view NMF for image annotation which learns a generative model specific to a particular query. The factorization is performed in such a way that ensures consistency in coefficients across features. This yields an elegant solution to the problem of feature fusion. Furthermore, we introduce weight matrices which increase the recall of the rare tags, without requiring tag-specific discriminative models. The proposed solution is practical for real world datasets characterized by continuous addition of images and tags.

2. Related Work

Over the past decade, significant efforts have been devoted to the task of image annotation. Many approaches are generative in nature consisting of either the topic or mixture models. In mixture model-based approaches, each annotated image is modeled as a mixture of topics over visual and tag features, where the mixture proportions are shared between different features or views. Examples include latent Dirichlet allocation [1], probabilistic latent semantic analysis [24], hierarchical Dirichlet processes [35], machine translation methods [6], and canonical correlation analysis [27]. The approach by Xiang et al. [32] also performs query-specific training using Markov random fields

but it has expensive testing as one MRF is generated per tag. Mixture models define a joint distribution over image features and annotations. Given a query image, these models compute the conditional probability over tags given the visual features by marginalizing the joint likelihood. Carneiro et al. [2] use a fixed number of mixture components over visual features per tag, while in [7], it is defined by using the training images as components over visual features and tags. Yavlinsky et al. [37] annotate images using only global features and perform nonparametric density estimation over the features. Besides generative approaches, discriminative models have also been used including SVM [4, 30], ranking SVM by Grangier et al. [11] and the method by Hertz et al. [16] which uses boosting.

A number of recent papers have reported better results with simple data-driven approaches by finding visually similar training images for a given query followed by transfer of tags from those images. Joint Equal Contribution (JEC) by Makadia et al. [22] was one of the first papers to highlight the effectiveness of the nearest-neighbors (NN) for image annotation. The paper presented an ad-hoc but simple procedure to transfer annotations from NN to the query image. The authors found that equal contributions from different features (mean of distances) performs on par with computationally expensive L_1-regularized Logistic Regression (Lasso). In contrast, we propose to fuse features using Multi-view NMF and show that it improves results. Guillaumin et al. [15] introduced TagProp which also uses nearest-neighbors to transfer tags. They showed that using large number of features, metric learning and special handling of rare tags (tag-specific models) improve results of image annotation. The nearest-neighbors are employed both during training and testing. Verma and Jawahar [29] presented two-pass kNN to find neighbors in semantic neighborhoods besides metric learning which learns weights for combining different features. The nearest-neighbor search they require for their method scales superlinearly with the number of training images, as a single image can occur in multiple semantic neighborhoods.

Non-negative matrix factorization has been successfully applied to various domains including text (document clustering [33]) and vision (face recognition [14]) and, in general, is an active area of research in clustering. Unlike PCA, the non-negativity of coefficients can be readily translated as weighted-assignment to basis or clusters. To understand the relationship between PCA, VQ and single-view NMF, the reader is referred to [18, 14]. The work by [12] proposes to integrate multiple views but the optimization is not performed jointly among views. They also propose a model selection strategy for identifying the correct number of clusters (basis). The work by Liu et al. [21] proposes a multi-view extension of NMF along with a novel normalization which makes all the basis to have unit sum permitting interpretation in terms of pLSA [10, 5]. Our approach is a weighted extension of [21], and differs in three aspects. First, [21] uses Multi-view NMF for unsupervised data clustering, i.e. they assume all data is available. For the task of image annotation which is supervised multi-label concept detection, testing data is not known. Thus, we need to recover coefficients for the query image during testing. Second, in our our formulation, we introduce weight matrices to handle imbalanced data. The third and most important difference is that, while [21] only uses visual features, we use tags as another feature and force Multi-view NMF to learn a set of basis across visual features and tags that are consistent across views. The key insight is that if we learn basis across features enforcing consistency on coefficients, then it is possible to use learned tag basis and query's visual features to obtain tags.

The proposed use of Multi-view NMF is also related to Relaxed Collaborative Representation [36], but rather than using features of training images directly as dictionaries, we learn a query-specific set of basis in each view and use that to transfer tags from annotated nearest-neighbors to the query image. One can also see our proposed approach as a multi-view extension with multiple weight matrices of the weighted but single-view NMF [13].

3. Proposed Approach

Given a query image, we first find its nearest-neighbors in the database which are assumed to be annotated with tags. Each image is represented in terms of visual features which we treat as different views of the image. We also treat tags as another view by obtaining binary vectors with length equal to the vocabulary size of tags. Then, the matrices from all views are decomposed to obtain basis in each view such that the coefficient vector of each NN image is consistent across all views. This gives a query-specific generative model from which the tags of query image are generated using its visual features.

3.1. Weighted Multi-view Non-negative Matrix Factorization (Query-specific Training)

Given a query image represented with multiple visual features, we compute its distance to images in the database in each view using pre-defined distance metrics (see Sec. 4). Next, the distance is normalized to lie between 0 and 1 for all the images for each view. Then, the distances across views are combined by taking their average and the N images with the smallest average distance are selected as nearest-neighbors.

Let $\mathbf{X}^{(f)} \in \mathbb{R}^{M_f \times N}$ represent the matrix obtained by horizontal concatenation of features vectors of length M_f from N images in f-th view, with a total of F views. Since we treat tags as another view, we let $\mathbf{X}^{(F+1)} = \mathbf{X}^{(tag)}$.

The goal of Multi-view NMF is to decompose each $\mathbf{X}^{(f)}$ into a basis matrix $\mathbf{U}^{(f)} \in \mathbb{R}^{M_f \times K}$ and a coefficient matrix $\mathbf{V}^{(f)} \in \mathbb{R}^{N \times K}$, where the parameter K defines the number of basis or latent concepts in each view. The factorization is subjected to soft-consensus regularization which enforces coefficient vectors corresponding to each image to be similar to a consensus vector in all views. This also results in the basis vectors to capture similar contents in their respective views. This is particularly desirable for image annotation as the coefficient vectors, recovered using the basis and the visual features of the query, are comparable across views.

Furthermore, to improve predictability of rare tags, we introduce two weight matrices in Multi-view NMF formulation which bias the factorization towards improved reconstruction for rare tags. Weight matrix $\mathbf{T}^{(f)} \in \mathbb{R}^{M_{(f)} \times M_{(f)}}$ is identity for views corresponding to visual features. However, it is a diagonal matrix for tag-view and is used to increase weight of rare tags so that reconstruction is biased towards a solution which results in improved performance on such tags. The matrix $\mathbf{W} \in \mathbb{R}^{N \times N}$ gives more weight to images containing rare tags and is applied to all views.

The objective function for Weighted Multi-view NMF which constitutes reconstruction and regularization terms is given by:

$$\begin{aligned} L = &\sum_{f=1}^{F+1} \|\mathbf{T}\big(\mathbf{X}^{(f)} - \mathbf{U}^{(f)}\mathbf{V}'^{(f)}\big)\mathbf{W}\|_F^2 \\ &+ \sum_{f=1}^{F+1} \lambda_f \|\mathbf{W}'\big(\mathbf{V}^{(f)}\mathbf{Q}^{(f)} - \mathbf{V}^*\big)\|_F^2 \\ &\text{s.t. } \forall 1 \le f \le F+1, \quad \mathbf{U}^{(f)}, \mathbf{V}^{(f)}, \mathbf{V}^* \ge 0, \end{aligned} \tag{1}$$

where $(.)'$ denotes transpose operator. Since the factorization obtained by NMF is not unique, i.e., $\mathbf{U}\mathbf{V}' = \mathbf{U}\mathbf{Q}^{-1}\mathbf{Q}\mathbf{V}'$, we use the $\mathbf{Q}$ to normalize $\mathbf{U}$ so that each basis vector sums to 1, i.e., $\|\mathbf{U}_{:,i}\| = 1$. The diagonal matrix $\mathbf{Q}^{(f)} \in \mathbb{R}^{K \times K}$ is defined as:

$$\mathbf{Q}^{(f)} = \text{Diag}\Big(\sum_{m=1}^{M} \mathbf{U}_{m,1}^{(f)}, \sum_{m=1}^{M} \mathbf{U}_{m,2}^{(f)} \dots \sum_{m=1}^{M} \mathbf{U}_{m,K}^{(f)}\Big). \tag{2}$$

The minimization is performed through an iterative procedure. First, we minimize L over $\mathbf{U}^{(f)}$ and $\mathbf{V}^{(f)}$ keeping $\mathbf{V}^*$ fixed. In the next step, we minimize L over $\mathbf{V}^*$ keeping $\mathbf{U}^{(f)}$ and $\mathbf{V}^{(f)}$ fixed. The procedure is repeated for a fixed number of iterations. Both $\mathbf{U}^{(f)}$ and $\mathbf{V}^{(f)}$ are initialized with non-negative values. Since the function is non-convex, optimization converges to a local minima.

Minimize over $\mathbf{U}^{(f)}$ and $\mathbf{V}^{(f)}$, given $\mathbf{V}^*$: With $\mathbf{V}^*$ fixed, $\mathbf{U}^{(f)}$ does not depend on $\mathbf{U}^{(f')}$ and $\mathbf{V}^{(f')}$ for $f' \neq f$. In the following treatment, we drop notation of feature for clarity. The objective function for a particular feature for fixed $\mathbf{V}^*$ is given by,

$$\begin{aligned} &\|\mathbf{T}\big(\mathbf{X} - \mathbf{U}\mathbf{V}'\big)\mathbf{W}\|_F^2 + \lambda_f \|\mathbf{W}'\big(\mathbf{V}\mathbf{Q} - \mathbf{V}^*\big)\|_F^2 \\ &\text{s.t. } \mathbf{U}, \mathbf{V} \ge 0. \end{aligned} \tag{3}$$

Compute $\mathbf{U}^{(f)}$, given $\mathbf{V}^{(f)}$ and $\mathbf{V}^*$: We obtain the following multiplicative update rule (similar to [21]):

$$\mathbf{U}_{i,k} \leftarrow \mathbf{U}_{i,k} \nabla_{\mathbf{U}_{i,k}} \text{ and } \nabla_{\mathbf{U}_{i,k}} = \frac{(\mathbf{T}'\mathbf{T}\mathbf{X}\mathbf{W}\mathbf{W}'\mathbf{V})_{i,k} + \lambda_f \sum_{n=1}^{N} \mathbf{W}_{n,n}^2 \mathbf{V}_{n,k} \mathbf{V}_{n,k}^*}{(\mathbf{T}'\mathbf{T}\mathbf{U}\mathbf{V}'\mathbf{W}\mathbf{W}'\mathbf{V})_{i,k} + \lambda_f \sum_{m=1}^{M} \mathbf{U}_{m,k} \sum_{n=1}^{N} \mathbf{W}_{n,n}^2 \mathbf{V}_{n,k}^2} \tag{4}$$

Compute $\mathbf{V}^{(f)}$, given $\mathbf{U}^{(f)}$ and $\mathbf{V}^*$: We obtain the following multiplicative update rule (similar to [21]):

$$\mathbf{V}_{j,k} \leftarrow \mathbf{V}_{j,k} \nabla_{\mathbf{V}_{j,k}} \text{ and } \nabla_{\mathbf{V}_{j,k}} = \frac{(\mathbf{W}\mathbf{W}'\mathbf{X}'\mathbf{T}'\mathbf{T}\mathbf{U})_{j,k} + \lambda_f \mathbf{W}_{j,j}^2 \mathbf{V}_{j,k}^*}{(\mathbf{W}\mathbf{W}'\mathbf{V}\mathbf{U}'\mathbf{T}'\mathbf{T}\mathbf{U})_{i,k} + \lambda_f \mathbf{W}_{j,j}^2 \mathbf{V}_{j,k}} \tag{5}$$

Minimize over $\mathbf{V}^*$, given $\mathbf{U}^{(f)}$ and $\mathbf{V}^{(f)}$: Once $\mathbf{U}^{(f)}$ and $\mathbf{V}^{(f)}$ have been updated for each view in a particular iteration, we take derivative of (1) w.r.t $\mathbf{V}^*$, set it equal to 0 and obtain the following closed-form solution to $\mathbf{V}^*$:

$$\mathbf{V}^* = \frac{\sum_{f=1}^{F} \lambda_f \mathbf{W}\mathbf{W}'\mathbf{V}^{(f)}\mathbf{Q}^{(f)}}{\sum_{f=1}^{F} \lambda_f \mathbf{W}\mathbf{W}'} \tag{6}$$

3.2. Boosting Mechanism for Rare Tags

Since rare tags appear with low frequency, they are overshadowed by frequent tags during training which leads to low recall for rare tags. The matrices $\mathbf{T}^{(tag)}$ and $\mathbf{W}$ introduced in the NMF improve recall and address the issue of dataset imbalance. Weight matrix $\mathbf{T}^{(tag)}$ is a diagonal matrix with $\mathbf{T}_{i,i}^{(tag)}$ set to 1/frequency of i-th tag in a query's neighborhood. $\mathbf{T}^{(tag)}$ penalizes inaccurate matrix factorization in Eq.1 severely for rare tags to ensure that the learned $\mathbf{U}^{(tag)}$ accurately models the rare tags. Thus, all tags contribute equally to the loss function. Furthermore, the diagonal matrix $\mathbf{W}$ is embedded in Eq.1 to bias the learned basis matrices towards a more accurate factorization of images with rare tags. $\mathbf{W}_{j,j}$ equals the summation of 1/frequency of tags of j-th NN example. The images annotated with rare tags are more important for an accurate generative model around a query as they capture co-occurrence of rare tags with the more frequent ones.

3.3. Recovering Tags of Query (Testing)

Given the factorized basis using features and tags of nearest-neighbors of a particular query image, we recover

the coefficients for visual features of the query in terms of learned basis using GPSR[8], which gives stable results than least squares especially when basis matrix is ill-conditioned. Next, the coefficient vectors from all views are combined using weighted average with weight for view f equal to λ_f to estimate the coefficient vector $\tilde{\mathbf{V}}^{(tag)}$. In our experiments, those were set to 0.01 for visual features and 1 for tags, so that basis for visual features are aligned with those of tags. Finally, the product of $\tilde{\mathbf{V}}^{(tag)}$ with the $\mathbf{U}^{(tag)}$ which was learned during training, i.e., $\mathbf{U}^{(tag)}(\tilde{\mathbf{V}}^{(tag)})'$ gives the scores for predicted tags. The desired number of tags can be obtained by ranking the tags according to the obtained scores.

4. Experiments

In this section, we first explain the datasets used in our experiments as well as metrics used for evaluation. Next, we present the experimental results of proposed method on different datasets and compare them to previous works. Finally, we evaluate the effect of weight matrices and conclude the section with a brief discussion on the complexity of the proposed method.

4.1. Datasets and Evaluation Metrics

We performed experiments on two popular and publicly available datasets Corel5K and ESP Game. Initially used by [6], Corel5K is the most common dataset for tag-based image annotation and retrieval. The training and testing sets consist of $4,500$ and 499 images, respectively. Images are manually annotated with, 3.4 tags on average, from a dictionary of 260 tags. ESP Game contains images annotated through an on-line game [31] in which players had to predict the same tags for images to gain points. Training and testing sets of ESP Game contain $18,689$ and $2,081$ images, respectively, with each image has 4.7 tags on average, from a dictionary of 268 tags. Figure 2 shows some example images from ESP Game dataset where we also illustrate the tag ambiguity for two images which share many tags while being visually and conceptually dissimilar.

We follow the evaluation metrics used in [15]. We automatically annotate each image with 5 tags and then compute precision and recall for each tag. The average precision (P) and average recall (R) across all tags in addition to the number of tags with non-zero recall ($N+$) is reported for performance evaluation. The F_1 measure, defined as harmonic mean of P and R ($F_1 = 2\frac{P*R}{P+R}$), is also reported.

4.2. Features

We used the publicly available features[4] provided by [15]. These are categorized as global and local descriptors. Global descriptors consist of GIST [26] and color histograms of RGB, Lab and HSV. Local descriptors include SIFT and robust (invariant to lighting geometry and specularities) hue descriptor [28] extracted around multi-scale grid and Harris-Laplacian interest points. Color histograms, SIFT and hue descriptors are also computed over three equal horizontal partitions (denoted by V3H1) for each image to encode spatial information. This provides a total of $F = 15$ features representing each image. To compare two features, we used L_1 for color histograms, L_2 for GIST and χ^2 for SIFT and hue descriptors, as was done in [15].

(a)

(b) circle, orange, round (c) circle, music, orange, red, round, white

Figure 2. Example images from ESP Game dataset are illustrated in 2(a). Figures 2(b) and 2(c) share many tags, although they are conceptually and visually different.

4.3. Results

Table 1 compares the performance of the proposed NMF-KNN framework to existing approaches on Corel5k dataset. We can see that NMF-KNN significantly outperforms other image annotation algorithms including the ML variant of TagProp which does not use tag-specific discriminative models. This indicates that the proposed approach is more effective than weighted nearest-neighbor based approaches. To handle the issue of rare tags and boost their recall, TagProp [15] learns discriminant models for each tag given by the variant TagProp-σML, which is the state-of-

[4]Features: GIST, DenseSIFT, DenseSIFTV3H1, HarrisSIFT, HarrisSIFTV3H1, DenseHue, DenseHueV3H1, HarrisHue, HarrisHueV3H1, RGB, RGBV3H1, Lab, LabV3H1, HSV, HSVV3H1 - available at http://tinyurl.com/l5d68sj

Figure 3. Example images from ESP Game dataset and the corresponding top 5 tags predicted using NMF-KNN are shown in this figure. Predicted tags in green appear in the ground truth while red ones do not. In many cases, even though the proposed method has predicted relevant tags to the image, those tags are missing in the ground truth. That is because the tag lists are not complete and are generally a subset of relevant tags.

Method	P	R	F_1	$N+$
CRM[17]	16	19	17.3	107
InfNet[23]	17	24	19.9	112
NPDE[37]	18	21	19.3	114
SML[2]	23	29	25.6	137
MBRM[7]	24	25	24.4	122
TGLM[20]	25	29	26.8	131
JEC[22]	27	32	29.2	139
TagProp-ML[15]	31	37	33.7	146
TagProp-σML[15]	33	42	36.9	160
Group Sparsity[38]	30	33	31.4	146
FastTag[3]	32	43	36.7	166
NMF-KNN	38	56	45.2	150

Table 1. Performance evaluation on Corel5k dataset

Method	P	R	F_1	$N+$
MBRM[7]	18	19	18.4	209
JEC[22]	22	25	23.4	224
TagProp-σSD[15]	39	24	29.7	232
TagProp-ML[15]	49	20	28.4	213
TagProp-σML[15]	39	27	31.9	239
FastTag[3]	46	22	29.7	247
NMF-KNN	33	26	29.0	238

Table 2. Performance evaluation on ESP Game dataset

the-art algorithm on this dataset. Although, TagProp-σML has a slightly higher $N+$ with a difference of 10, the proposed method gives a much higher P, R, and F_1 making it competitive to TagProp-σML.

Table 2 shows the results of the proposed and comparison methods on ESP Game dataset. The proposed method provides competitive results w.r.t R and $N+$ to TagProp-σML. This shows that learning a model around a query is more useful and the natural capabilities of features fusion and handling rare tags, without requiring training on entire datasets, makes our approach superior to the previous methods. NMF-KNN performs slightly worse than FastTag [3], however, at constant time complexity during testing. Figure 3 illustrates qualitative results of image annotation using proposed method on some example images from ESP Game dataset. True positives, i.e., the tags predicted by our method that also occur in ground truth are shown in green, while false positives are shown in red. It is evident that many of the predicted tags are relevant to the image content, even though, they are not annotated in the ground truth. Another important difference between our method and existing methods [22, 15, 29] is the number of nearest-neighbors used to propagate the tags. These methods retrieve around $N = 200$ neighbors per query while we use only $N = 40$ neighbors. This suggests that NMF-KNN can build a reliable model around the query with 20% data compared to the competitive methods.

To measure the effect of K, we evaluated the performance of NMF-KNN by increasing the value of K from 10 to 150 when N is fixed to 40. We observed that, for K beyond 50, R does not improve significantly while P initially increases and then reaches a plateau. Meanwhile, a larger K increases the computational complexity of the model and therefore is not desirable. We also studied the effect of neighborhood size, N, on the performance of our proposed method. For N larger than 40, we did not observe a considerable change in R, however P begins to decrease. A possible explanation is that for large neighborhood sizes,

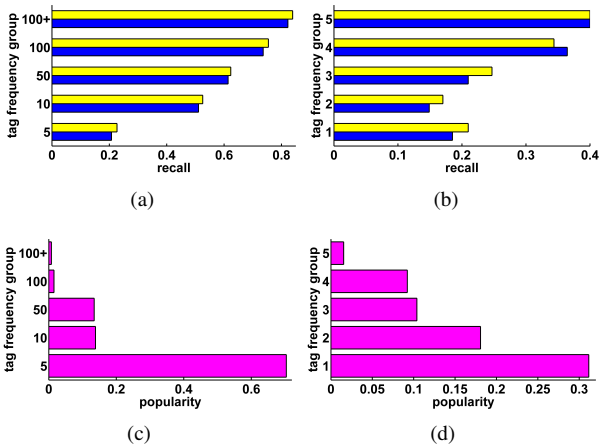

Figure 4. Evaluating the effect of Weight Matrices: Evaluated on Corel5k dataset, 4(a) shows the effect of using weight matrices, before (blue) and after (yellow), on the annotation performance. Tags are grouped based on their frequency of appearance in the dataset. The first bin groups words that have between 1 to 5 images related to them. The second bin is associated with tags with the images between 6 to 10, and so on. In 4(b) we show the same for first group of 4(a) to analyze the recall of tags with 1 to 5 related images. 4(c) and 4(d) give the fraction of tags in each bin of 4(a) and 4(b), respectively. This shows that we can improve the recall of rare tags without sacrificing that of frequent tags.

we allow irrelevant training examples to participate in construction of our query-specific model and therefore, learned basis become contaminated.

4.4. Effect of W and T

To evaluate the proposed boosting mechanism for rare tags, we study the effect of $\mathbf{W}$ and $\mathbf{T}$ in the Multi-view NMF. In Fig. 4(a) and 4(b), the y-axis shows the frequency with which tags appear in the training dataset. In Fig. 4(a), they have been grouped, while in Fig. 4(b), we show them individually. The x-axis shows the value of recall and the blue and yellow bars represent the before and after effect of $\mathbf{W}$ and $\mathbf{T}$, respectively. In Fig. 4(c) and 4(d), the fraction of tags that belong to each group are shown. Boosting mechanism improves the mean recall of tags in five groups (Fig. 4(a)) by $\mathbf{1.91\%}$, $\mathbf{1.48\%}$, $\mathbf{0.94\%}$, $\mathbf{1.82\%}$ and $\mathbf{1.67\%}$, respectively. The first group contains tags with frequency less than 6 while the last one with frequency greater than 100. Fig. 4(c) shows that the majority ($\mathbf{70.38\%}$) of tags are assigned to less than 6 images in the dataset. From Fig. 4(d), we can see that tags with only 1 relevant image in the dataset are dominant. The proposed boosting mechanism increased the recall of tags with 1, 2 and 3 relevant tags in the dataset by $\mathbf{2.47\%}$, $\mathbf{2.13\%}$ and $\mathbf{3.70\%}$, respectively. In summary, Fig. 4(a) shows that for the tags belonging to all frequencies, the boosting mechanism improves the recall. This is different from TagProp [15] which sacrifices recall of frequent tags to boost that of rare tags.

4.5. Computational Complexity

As noted by [9, 3, 38], [15]'s training complexity is quadratic, $O(n^2)$, where n is the number of training images. Since it relies on sophisticated training procedures and per tag optimizations, it is not scalable on large datasets. Adding new images or tags to the dataset influences the performance of trained models as both positive and negative instances change for discriminative classifiers. JEC [22] and FastTag [3] are comparable with proposed method in terms of complexity but [22] provides considerably lower performance. Since [3] performs a global co-regularized learning, regressor (W) and enricher (B) matrices have to be retrained when a new set of samples or tags are introduced to the dataset.

The proposed Multi-view NMF framework does not require any training but has $O(n)$ test-time complexity due to nearest-neighbor look up for the query image where n is the total number of training examples. The complexity of Weighted Multi-view NMF is linear with respect to the cardinality of chosen nearest-neighborhood that results in $O(n)$ complexity for the proposed approach.

In our experiments, query-specific training usually converges after $15-20$ iterations. The computation cost breakdown of NMF-KNN follows: 80% for finding the nearest-neighbors, 19% for learning the model and 1% for predicting tags. Sub-linear time complexity can be achieved by employing approximate NN search methods e.g. FLANN [25] or k-d trees in the implementation of NMF-KNN.

5. Conclusion and Future Work

The proposed approach is suitable to real-world databases which are characterized by a continuous increase in both the images and tags assigned to those images. It is a hybrid of nearest-neighbor and generative approaches and does not require any training in the form of global mapping between features and tags or tag-specific discriminant models. NMF-KNN allows feature-fusion in a mathematically coherent way by discovering a set of basis using both the visual features and annotated tags. The weight matrices handle the issue of dataset imbalance by increasing the recall of rare tags. The proposed approach offers a practical solution to real world datasets and performs competitively with state-of-the-art methods without requiring any training. This is due to the proposed Weighted Multi-view NMF which learns a superior model specific to each query while requiring fewer number of nearest-neighbors for tag transfer. For future work, we intend to find the weight matrices within optimization instead of pre-defining them. We would also like to explore the possibility of introducing kernels within the proposed framework.

Acknowledgment This work was supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract numbers D11PC20066. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S.Government.

References

[1] K. Barnard, P. Duygulu, D. Forsyth, N. De Freitas, D. M. Blei, and M. I. Jordan. Matching words and pictures. *JMLR*, 3(2), 2003.

[2] G. Carneiro, A. B. Chan, P. J. Moreno, and N. Vasconcelos. Supervised learning of semantic classes for image annotation and retrieval. *PAMI*, 29(3), 2007.

[3] M. Chen, A. Zheng, and K. Weinberger. Fast image tagging. In *ICML*, 2013.

[4] C. Cusano, G. Ciocca, and R. Schettini. Image annotation using SVM. In *Electronic Imaging*, 2003.

[5] C. Ding, T. Li, and W. Peng. On the equivalence between non-negative matrix factorization and probabilistic latent semantic indexing. *Computational Statistics & Data Analysis*, 52(8), 2008.

[6] P. Duygulu, K. Barnard, J. F. de Freitas, and D. A. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *ECCV*, 2002.

[7] S. Feng, R. Manmatha, and V. Lavrenko. Multiple bernoulli relevance models for image and video annotation. In *CVPR*, 2004.

[8] M. A. Figueiredo, R. D. Nowak, and S. J. Wright. Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems. *IEEE Journal of Selected Topics in Signal Processing*, 1(4), 2007.

[9] H. Fu, Q. Zhang, and G. Qiu. Random forest for image annotation. In *ECCV*. 2012.

[10] E. Gaussier and C. Goutte. Relation between pLSA and NMF and implications. In *ACM SIGIR Research and Development in Information Retrieval*, 2005.

[11] D. Grangier and S. Bengio. A discriminative kernel-based approach to rank images from text queries. *PAMI*, 30(8), 2008.

[12] D. Greene and P. Cunningham. A matrix factorization approach for integrating multiple data views. In *Machine Learning and Knowledge Discovery in Databases*. 2009.

[13] D. Guillamet, M. Bressan, and J. Vitria. A weighted non-negative matrix factorization for local representations. In *CVPR*, 2001.

[14] D. Guillamet, B. Schiele, and J. Vitria. Analyzing non-negative matrix factorization for image classification. In *CVPR*, 2002.

[15] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid. Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In *ICCV*, 2009.

[16] T. Hertz, A. Bar-Hillel, and D. Weinshall. Learning distance functions for image retrieval. In *CVPR*, 2004.

[17] V. Lavrenko, R. Manmatha, and J. Jeon. A model for learning the semantics of pictures. In *NIPS*, 2003.

[18] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755), 1999.

[19] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *NIPS*, 2000.

[20] J. Liu, M. Li, Q. Liu, H. Lu, and S. Ma. Image annotation via graph learning. *Pattern Recognition Journal*, 42(2), 2009.

[21] J. Liu, C. Wang, J. Gao, and J. Han. Multi-view clustering via joint nonnegative matrix factorization. In *SIAM Data Mining Conf.*, 2013.

[22] A. Makadia, V. Pavlovic, and S. Kumar. A new baseline for image annotation. In *ECCV*, 2008.

[23] D. Metzler and R. Manmatha. An inference network approach to image retrieval. In *Image and Video Retrieval*. 2004.

[24] F. Monay and D. Gatica-Perez. pLSA-based image auto-annotation: constraining the latent space. In *ACM MM*, 2004.

[25] M. Muja and D. G. Lowe. Fast matching of binary features. In *Computer and Robot Vision*, 2012.

[26] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, 42(3), 2001.

[27] R. Socher and L. Fei-Fei. Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora. In *CVPR*, 2010.

[28] J. Van De Weijer and C. Schmid. Coloring local feature extraction. In *ECCV*. 2006.

[29] Y. Verma and C. Jawahar. Image annotation using metric learning in semantic neighbourhoods. In *ECCV*. 2012.

[30] Y. Verma and C. Jawahar. Exploring svm for image annotation in presence of confusing labels. In *BMVC*, 2013.

[31] L. Von Ahn and L. Dabbish. Labeling images with a computer game. In *SIGCHI Conference on Human Factors in Computing Systems*, 2004.

[32] Y. Xiang, X. Zhou, T.-S. Chua, and C.-W. Ngo. A revisit of generative model for automatic image annotation using markov random fields. In *CVPR*, 2009.

[33] W. Xu, X. Liu, and Y. Gong. Document clustering based on non-negative matrix factorization. In *ACM SIGIR Research and Development in Informaion Retrieval*, 2003.

[34] X. Xu, A. Shimada, and R.-i. Taniguchi. Image annotation by learning label-specific distance metrics. In *ICIAP*. 2013.

[35] O. Yakhnenko and V. Honavar. Annotating images and image objects using a hierarchical dirichlet process model. In *ACM Int. W. on Multimedia Data Mining*, 2008.

[36] M. Yang, L. Zhang, D. Zhang, and S. Wang. Relaxed collaborative representation for pattern classification. In *CVPR*, 2012.

[37] A. Yavlinsky, E. Schofield, and S. Rüger. Automated image annotation using global features and robust nonparametric density estimation. In *Image and Video Retrieval*, 2005.

[38] S. Zhang, J. Huang, Y. Huang, Y. Yu, H. Li, and D. N. Metaxas. Automatic image annotation using group sparsity. In *CVPR*, 2010.