IMAGE PATTERN DISCOVERY BY USING THE SPATIAL CLOSENESS OF VISUAL CODE WORDS

Meng Sun, Hugo Van hamme

Department of Electrical Engineering-ESAT, Katholieke Universiteit Leuven,
Kasteelpark Arenberg 10, Bus 2441, B-3001 Leuven, Belgium

ABSTRACT

A graph regularized non-negative matrix factorization (NMF) model is proposed for image pattern discovery. Each image is represented by its histogram of visual words (i.e. bag-of-words) and the image contents are discovered by the NMF model. The graph regularization preserves the spatial closeness of visual code words in the obtained patterns, thus improving the bag-of-words representation against its main shortcoming: the loss of spatial information. Experiments on a subset of the Caltech256 database show the efficacy of the proposed model.

Index Terms— Image representation, non-negative matrix factorization, spatial information, graph regularization

1. INTRODUCTION

Image pattern discovery is a hot topic in computer vision research. Usually its performance depends a lot on the amount of annotated training data, which requires huge human efforts. Therefore, a number of unsupervised models were proposed to discover useful patterns from images: k -means clustering, probabilistic latent semantic analysis (pLSA), non-negative matrix factorization (NMF), kernel principle component analysis (KPCA), etc.[1]. The k -means method yields a good partitioning of the training data space by maximizing the total inter-class distance. KPCA compares the similarity of images by projecting them onto a new space. pLSA and NMF are good at extracting recurring parts in images, making them plausible choices for the unsupervised pattern discovery task.

The *bag-of-words* representation is popular for images, where the *words* are clustering centers of the vectors of scale invariant feature transforms (SIFT) [2]. The *bag-of-words* is additive, that is the representation of an image approximates the sum of the representations of its *parts*. However, the spatial information is lost when transforming an image to its *bag-of-words*, thus two different patterns could end up with similar representations, which limits the classification/recognition ability of the model. In the supervised models, this ambiguity between patterns can be avoided as the model has to fit the

ground truth information during training. However, without supervision, the model can find any solution which locally minimizes the cost function given the training data, while the solutions can be physically meaningless [3].

In this paper, we impose spatial constraints on the visual words to refine the NMF model and obtain better results in unsupervised training. The motivation for this is two fold. One motivation is to unify spectral clustering [4] and latent variable analysis (NMF). Spectral clustering methods are good at **agglomerating** data points by their proximities, thus working in a *bottom-up* way; NMF **decomposes** objects into recurring patterns which are contained in the objects, hence working *top-down*. By incorporating the two, the model finds recurring patterns as well as clustering the features into patterns, so better performance can be expected. The other one is motivated by the difficulty to determine a proper codebook size. Small codebooks provide only inaccurate descriptors of the salient features of the patterns, which are important for the training of the NMF model. Too large a codebook will make the activations of code words very sparse, thus the recurrence frequencies of code words would be quite low, which would make the NMF model worse as the model can only find *recurring* patterns.

The structure of the paper is as follows. The models and algorithms are explained in Section 2. The results are given in Section 3. The discussion and conclusion are in Section 4 and Section 5 respectively.

2. GRAPH REGULARIZED NMF MODEL FOR IMAGE OBJECT DISCOVERY

2.1. Representation of images: bag of SIFT code words

SIFT descriptors can be extracted for a gray scale image [2]. A descriptor has two kinds of information. The first one is the location $e_t = (x_t, y_t, r_t, o_t)$, where (x_t, y_t) is the coordinate of the interest point, r_t is the scale at which the interest point was extracted and o_t is the main direction of the change of gray scales at the interest point. We discard the orientation o_t in this paper. The second one is a histogram d_t with 4×4 directions each of which has 8 bins of gray scales. For each interest point, d_t is a 128 dimensional vector.

The research was funded by the K.U.Leuven research grant OT/09/028(VASI).

By clustering a part of or all the descriptors $\{d_t\}$ of the training images using k -means, we obtain M cluster centers, called visual words. The descriptors of training and test set are subsequently vector quantized (VQ) by the visual words. Two VQ techniques are tried in this report: hard-VQ and soft-VQ. In hard-VQ or *winner-takes-all*, a descriptor is given the label of the visual word at minimal Euclidean distance. In soft-VQ, the top K visual words and their associated scores are selected for a descriptor. Suppose the first $K + 1$ Euclidean distances (sorted in increasing order) are $z_{i_1}, z_{i_2}, \dots, z_{i_{K+1}}$, then the score on the i_k -th visual word is defined as $e^{-z_{i_k}/z_{i_{K+1}}}$. Then a transformation is made to scale all scores between 0 and 1 in every retained code word.

$$p(f_{i_k}; e_t) = \frac{e^{-\frac{z_{i_k}}{z_{i_{K+1}}}} - e^{-1}}{1 - e^{-1}}, 1 \leq k \leq K + 1 \quad (1)$$

where f_{i_k} is the i_k -th code word and e_t is the interest point. The $K + 1$ -th score is always zero, so the top K visual words are actually retained for the descriptor e_t . By normalizing the scores to sum to 1 over all code words, they can be interpreted as *posterior probabilities* of the descriptor e_t on the visual words: $p(f_{i_k}|e_t)$. By accumulating the posterior probabilities of all the descriptors of the image I as in Equation 2, an image is represented as a histogram of visual words.

$$\sum_{e_t \in I} p(f_k|e_t) \quad k = 1, \dots, M \quad (2)$$

For all the training images, the data matrix $V_{M \times N}$ is obtained, where N is the number of images in the training data and the element V_{ij} is the accumulated posterior probability on the i -th visual word of the j -th image.

2.2. Non-negative matrix factorization

In the training stage of the NMF model, given the data matrix $V_{M \times N}$ obtained in the last subsection, by solving $\min_{W, H} C_0(V||WH)$ where C_0 is the extended Kullback-Leibler divergence [5], we obtain the vocabulary patterns in the columns of $W_{M \times R}$ and the activations of the patterns in the columns of $H_{R \times N}$, where R is the number of patterns. Each column of W is normalized to sum to 1 to avoid trivial solutions and to retain the probabilistic rationale. The algorithms are described in [5] and will be the *unsupervised baseline* in this paper. The lack of supervisory data makes it however difficult to interpret the discovered patterns in W . For evaluation purposes, a mapping matrix Q is estimated to map the activations H (obtained with unsupervised learning) to the ground truth by solving $\min_Q C_0(G||QH)$, where G is constructed as follows: if the n -th image is known to belong to the l -th category, $G_{ln} = 1$; otherwise, $G_{mn} = 0$.

In the testing stage, we firstly estimate the activations H' of the previously trained patterns W by solving $\min_{H'} C_0(V'||WH')$, where V' is the data matrix for the test

set. Subsequently, the activations of the vocabulary items are computed as $\hat{G}' = QH'$ using the mapping matrix Q derived from the training data. $\hat{G}'$ indicates the category for each test image. The n -th image is assigned a category ID by taking the highest activated row number of the n -th column of $\hat{G}'$. Then classification error rates can be computed by comparing with the ground truth category information.

In [1], the *conditional entropy* is used to evaluate the performance of unsupervised clustering. The conditional entropy is based on the ground truth category labels X and obtained cluster labels Y of the test set, which means the average amount of uncertainty that remains about X once the value of Y is known. A mapping operator σ is used to map the obtained patterns to the ground truth categories, and the conditional entropy between X and $\sigma(Y)$ is computed. In this paper, Q is used to construct σ . The test image j is classified by $\arg\max_{kj} (QH')_{kj}$, where H' is the estimated activation coefficients of the text images. By comparing with the ground truth information, we obtain the estimate of the confusion matrix $C_{x\sigma(y)}$, where $C_{x\sigma(y)}$ is the number of images whose ground truth label is x while their predicted label is $\sigma(y)$. The conditional entropy is defined in Equation 3, where $p(x|\sigma(y)) = \frac{C_{x\sigma(y)}}{\sum_{\sigma(y)} C_{x\sigma(y)}}$ and $p(\sigma(y)) = \sum_x \frac{C_{x\sigma(y)}}{\sum_{x, \sigma(y)} C_{x\sigma(y)}}$.

$$H(X|\sigma(Y)) = \sum_{\sigma(y)} p(\sigma(y)) \sum_x p(x|\sigma(y)) \log \frac{1}{p(x|\sigma(y))} \quad (3)$$

2.3. Spatial closeness and graph regularization

In this part, we present how to group the spatially close visual words into one pattern by using an adjacency graph of visual words. Firstly, the spatial adjacency of two descriptors, (x_1, y_1, r_1, o_1) and (x_2, y_2, r_2, o_2) , is defined. If $\frac{r_1 + r_2}{\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}} > 0.9$, the two descriptors are said to be adjacent. A descriptor is also adjacent with itself. For each descriptor, the N_{best} visual words with the highest posterior probabilities are selected. The $N_{best} \times N_{best}$ visual words of two adjacent descriptors are adjacent. By accumulating the adjacency frequencies of the visual words over all the adjacent descriptors of the images of the training data, an adjacency matrix U is obtained, which is made symmetric by $U \leftarrow U + U^T$ and logical by $U \leftarrow U > \Gamma$, where Γ is a threshold. Next the diagonal elements of U are set to zeros. The degree matrix D is a diagonal matrix with diagonal elements equal to the column sum of U ; the graph Laplacian $L = D - U$.

The function of the graph regularization can be seen as in Equation (4) [4].

$$W_{:,k}^T L W_{:,k} = \sum_{u \sim v} (W_{uk} - W_{vk})^2 U_{uv} \quad (4)$$

where $W_{:,k}$ is the k -th column of W , $\sum_{u \sim v}$ denotes the sum

over all unordered features u, v for which u and v are adjacent in U . By minimizing Equation (4) for all the columns of W , the adjacent features (visual words) tend to be simultaneously highly activated or not in the same pattern $W_{:,k}$. Joined with the extended Kullback-Leibler divergence, the cost function of the optimization problem becomes Equation (5).

$$C(V||W, H) = C_0(V||W, H) + \lambda C_1(W) \quad (5)$$

where, $C_1(W) = \text{Tr}(W^T L W)$, λ is a regularization parameter to balance the two terms: divergence and regularization. In this way, we integrated the factorization model of images and the graph spectral clustering model of visual words to perform pattern discovery. Algorithms were proposed to solve the optimization problem in Equation 5 as well as retaining the normalization of columns of W [6].

Fig. 1. Results of image object discovery using different models. (a)(b): classification error rates with hard VQ and soft VQ respectively, (c)(d): the conditional entropy of hard VQ and soft VQ respectively. 1000, 2000 are sizes of codebooks.

3. RESULTS OF IMAGE OBJECT DISCOVERY

The models were evaluated on the Caltech 256 data set, limited to 20 categories as in [1]. For each selected category, 3/4 of the images were used for training. To avoid the influence of unbalanced sets, at most 80 images were selected for testing within the remaining 1/4 data¹. The extraction of SIFT features and the training of code words were made

¹The classification error rates of the unsupervised models on the unbalanced testing data (the 1/4 images) are around 35%.

with the software in [7] by using the default parameters. For soft VQ, the number of code words used for labeling a descriptor is $K=3$. The parameter to count adjacency of code words is $N_{best}=1$ in hard VQ and $N_{best}=3$ in soft VQ. The threshold to make the graph adjacency matrix U logical is $\Gamma \geq \frac{\# \text{training images}}{\# \text{Categories}} = \frac{2835}{20}$. The basic assumption for this is that an adjacency of some image category appears at least once in any image of the category. We took a slightly larger $\Gamma=200$ to avoid spurious adjacencies from noise.

The number of codebooks, the factorization dimension R and the regularization parameters λ were treated as adjustable parameters. The performance of the models with different parameters are in Figure 1. For the supervised baseline, we made the factorization $[G; V] = [Q; W]H$ to estimate Q and W and evaluated the performance as in Section 2.

4. DISCUSSION

4.1. Performance of the graph regularized NMF

With 1000 or 2000 code words, the graph regularized NMF always outperforms the corresponding unsupervised baseline. The improvement on the models with hard VQ is quite small, while the improvement on the models with soft VQ is obvious as shown in Figure 1. The reason is likely that the graph model requires sufficient adjacency between visual words to work, but the winner-takes-all technique in hard VQ discards too much useful connections between visual words. To increase our understanding, we now describe a method for visualizing graph adjacency. As is shown in (c)(d) of Figure 2, 2 images were taken for each of the two categories: category 18 (motorbikes) and category 7 (airplanes). For each image, we consider the adjacent interest points of the most central interest point (red circle). For the unsupervised model, the adjacency score of two descriptors e_1 and e_2 is: $\sum_{i=1}^M p(f_i|e_1)p(f_i|e_2)$, while for the graph regularized model, the score is $\sum_{i=1}^M p(f_i|e_1)p(f_i|e_2) + \sum_{i=1}^M \sum_{j=1, j \neq i}^M p(f_i|e_1)U_{ij}p(f_j|e_2)$, which also takes the proximity of different visual words into account. At most 10 adjacent descriptors (green circles) are shown in the images. The top ones are with unsupervised model, while the bottom ones are with graph adjacency. For the images of motorbike, we can see good connections among the descriptors on the object (motorbike) itself, while for the images of airplanes, a lot of background (e.g. ground, sky) is also connected with the descriptor on the plane.

4.2. Factorization dimension

Figure 1 shows that the model performs slightly better for larger R , where more complexity is available to model the variation in the images. Figure 2-(a) shows the mapping matrix Q of supervised learning with $R=30$. The model assigns different numbers of patterns to different categories: category

Fig. 2. (a)(b): Q between the obtained patterns (columns) and the ground truth categories (rows). (a):supervised soft VQ with $R=30$. (b):unsupervised soft VQ with $R=30$. (c)(d): interest points (green circles) that are connected to the interest point that is closest to the center of the image. Top: unsupervised NMF, bottom: graph regularized NMF.

7 and category 18 which have a lot of training examples were clustered into 4 resp. 2 sub-categories. However, with unsupervised models, the mapping matrix Q is not as good as in supervised cases (see (b) of Figure 2).

4.3. Hard VQ and soft VQ

Hard VQ always results in lower conditional entropy than soft VQ (compare (c) and (d) of Figure 1). However, there is significant improvement of the classification error rates from hard VQ to soft VQ in (a)(b) of Figure 1. A cause could be that by using soft VQ, we introduced a lot of uncertainty which increases the entropy. Notice that lower conditional entropy doesn't imply a lower error rate. The lowest entropy in [1] is 2.0 by using hard VQ and NMF models. Our results seem to be better than this baseline on this criterion, which is probably because we used a carefully obtained mapping matrix Q . Another advantage of soft VQ is that it makes the model robust. In our experiments, the standard deviation of the classification error rates of soft VQ is usually 1/3 of that of hard VQ. Supervised NMF with hard VQ yields better results than supervised NMF with soft VQ.

4.4. Large codebooks

Graph regularized NMF with a small number of code words (e.g. 1000) outperforms the unsupervised NMF with a larger number of code words (e.g. 2000) as seen in (b)(d) of Fig-

ure 1, which alleviates the difficulty of the selection the codebook size. But graph regularized NMF with too large codebooks doesn't improve the unsupervised baseline as seen from Table 1. The cause is that few code words ($\text{sum}(\text{diag}(D)) > 0$) in Table 1) can be regularized by the graph adjacency matrix when using very large codebooks. An improvement to this could be to construct the adjacency matrix in a non-data-driven way, e.g. by the geometric proximity of code words.

Table 1. Classification error rates w.r.t. codebook size, $R=25$

# code words	1000	2000	5000	10,000	20,000
supervised	40.60	37.98	36.36	37.17	33.86
unsupervised	53.46	52.26	49.76	46.99	48.40
graph reg. NMF	48.64	50.69	49.52	47.32	48.46
$\text{sum}(\text{diag}(D)) > 0$	1000	2000	2000	2000	1500

5. CONCLUSION

We presented a graph regularized NMF model for image pattern discovery to capture the spatial closeness of code words. The experiments on a subset of the Caltech256 database showed the efficacy of the proposed model. The model always outperforms the unsupervised baseline given medium codebook sizes. The effect of the VQ method (hard vs. soft) on two evaluation metrics (classification and conditional entropy) were also discussed.

6. REFERENCES

- [1] T.Tuytelaars, C.H.Lampert, M.B.Blaschko, W.Buntine, "Unsupervised Object Discovery: A comparison", *IJCV*, 88(2), pp.284-302, 2010.
- [2] D.G.Lowe, "Object recognition from local scale-invariant features", *ICCV*, pp.1150-1157, 1999.
- [3] D.Donoho, V.Stodden, "When does non-negative matrix factorization give a correct decomposition into parts?", *Seventeenth Annual Conference on Neural Information Processing Systems*, 2003.
- [4] F. Chung, "Spectral Graph Theory", *CBMS Regional Conference Series in Mathematics*, No.92, 1997.
- [5] D. Lee, H. Seung, "Algorithms for non-negative matrix factorization", *NIPS*, 2000.
- [6] M.Sun, H.Van hamme, "Unsupervised vocabulary discovery using non-negative matrix factorization with graph regularization", *ICASSP*, 2011.
- [7] <http://www.vlfeat.org/index.html>