Bag-of-Multimedia-Words for Image Classification

Amel ZNAIDIA, Aymen SHABOU, Hervé LE BORGNE
Laboratory of Vision and Content Engineering, CEA
{*amel.znaidia, aymen.shabou, herve.le-borgne*}*@cea.fr*

Céline HUDELOT, Nikos PARAGIOS
MAS Laboratory, Ecole Centrale de Paris
{*celine.hudelot, nikos.paragios*}*@ecp.fr*

Abstract

We introduce the bag-of-multimedia-words model that tightly combines the heterogeneous information coming from the text and the pixel-based information of a multimedia document. The proposed multimedia feature generation process is generic for any multimodality and aims at enriching a multimedia document description with compact and discriminative signatures well appropriate to linear classifiers. It is evaluated on the Pascal VOC 2007 classification challenge, outperforming the state-of-the-art bag-of-visual-words or bag-of-tag-words based classification approaches.

1. Introduction

Content-based document retrieval, as well as classification, implies an interest for document description, which precedes the learning or retrieval processes. In the case of multimedia documents, a key problem is the combination of different modalities that characterize a document. Two of the most commonly used modalities are those based on text and pixel-based information. For instance, in the case of annotated images, they correspond directly to the modalities that compose a document. In the case of videos, a large part of the information can be reduced to such modalities, via speech-to-text (for text) and keyframe extraction (for pixels).

Describing a document composed by both text and pixel-based information faces the problem of their heterogeneous nature. The textual modality is mapped into a dictionary that reflects a language or a sub-part of it into a particular domain, while the visual modality is usually transformed into feature vectors that form a low-level visual description. A common approach to tackle the problem of information's heterogeneity is to work around this issue by processing each modality separately and combining them at the decision level (late fusion). An alternative is to work on the description to make it more homogeneous (early fusion). A popular scheme in this vein is to use the *bag-of-word* (BoW) model, introduced in the text community [11]. In its simplest form, it consists in making an histogram of occurrences of words within a document (term counts). Many refinements have been proposed, such as taking into account the occurrence of words within the collection (inverse document frequency), the length of each document, and so on. This model has been introduced in the image community ten years ago [13], and its numerous extensions make it one of the most efficient representations used in image classification and retrieval. Words are then derived from local features such as SIFT [9], and the model is then named *bag-of-visual-words* (BoVW). Nevertheless, these descriptions do not directly convey human understandable meaning and a gap remains between them and the semantic content of images [14, 1].

In this paper, we propose a more integrated semantic signature for multimedia documents than the ones above, that results from a combination of textual and visual information. It is based on *multimedia codewords* that allow on the one hand cross-coding textual tag-words over visual-words extracted from a document; and on the other hand designing Bag-of-Multimedia-Words (BoMW) signatures. We exploit the recent advances in BoVW design methods [20, 2, 17, 8] in order to provide discriminative BoMW vectors suited to multimodal document classification with efficient linear classifiers.

The remainder of this paper is as follows. First the multimedia codeword is introduced in section 2. Then, designing multimedia signatures is detailed in section 3. Finally, some results are reported in section 4.

2. Multimedia codebook

2.1 Tag vs. Visual words

Tags provide contextual and semantic information which can be used to improve the accuracy of image classification [16, 4]. Such improvement however depends on the availability and quality of tags, which are often imprecise, ambiguous, overly personalized and in limited quantity for a given image [6, 12].

Visual words provide a low-level information to design BoVW signatures. However, the size of a visual-word vocabulary involves a trade-off between discriminatory power and computation cost. Indeed, with a small vocabulary, BoVW signature would be not discriminative enough because of assignment ambiguities of local features to codewords. As the size of the learned vocabulary increases, the signature becomes more discriminative, but meanwhile less generative and forgiving to noise, since similar descriptors would be mapped to different codewords. Furthermore, computational costs for designing BoVW signatures and classifying them grow. Currently, there is no consensus to the appropriate size of a visual vocabulary which varies from several hundreds [7], to tens of thousands [21] and even more.

To overcome the above problems, we define a multimedia word as the elementary part of a multimedia document similar to visual and tag words as elementary parts of an image and its corresponding caption.

2.2 Multimedia word

We denote by $\mathcal{T}_d$ the set of textual tags associated to a document d and $\mathcal{T}$ the set of all textual tags of the training dataset, with $\mathcal{T}_d \subset \mathcal{T}$ for each document d. A simple way to build the multimedia codebook is to perform two steps (1) tag-coding and (2) clustering.

The first steps consists in expressing each tag of $\mathcal{T}$ over a discrete visual codebook $\mathcal{W}^v$. This mapping, that we call *tag-coding*, relies on the fact that textual tag-words are semantically more consistent than visual-words, as it has also been observed in [10]. Therefore, coding tags over the visual codebook is much more interesting and coherent than the opposite way. Formally, let $\mathbf{V}$ be the visual word occurrence matrix learned on the training dataset $\mathcal{D}$ composed of N documents. $\mathbf{V}$ is of size $K^v \times N$, with K^v the size of a visual codebook $\mathcal{W}^v$. The tag-coding matrix $\mathbf{X}$ has the size $K^v \times K^t$, with K^t the size of the textual tag-codebook $\mathcal{W}^t$. To built $\mathbf{X}$, we sum for each textual tag the visual word occurrences across the list of images tagged with it, i.e.,

$$\mathbf{X}(i,j) = \sum_{d_k \in \mathcal{D},\, t_j \in \mathcal{T}_{d_k}} \mathbf{V}(i,k), \qquad (1)$$

with d_k the k^{th} document in the training dataset $\mathcal{D}$, t_j a tag in $\mathcal{T}_{d_k}$ and $\mathbf{V}(i,k)$ the occurrence of the i^{th} visual word in the document d_k. This matrix is then l_1 column normalized, expressing the frequency of a visual word relatively to a tag within the whole training dataset.

The second step, depicted in Figure 1, consists in clustering column vectors of $\mathbf{X}$, using K-means for instance, in order to generate the multimedia codebook (*M-codebook*), which is formed of relevant multimedia words. This step results in the following M-codebook :

$$\mathcal{W}^m = \left\{ \mathbf{m}_i \,;\, \mathbf{m}_i \in \Re^{K_v} \,;\, i = 1, ..., K^m \right\}, \qquad (2)$$

with K^m the size of the M-codebook ($K^m \leq K^t$).

3. Multimedia signature

From the obtained M-codebook, we generate a bag-of-multimedia-words (BoMW) signature. The process we propose takes advantage of advances in BoVW signature design [20, 2, 17, 8], that consists of two main steps *coding* and *pooling* as shown in Figure 1.

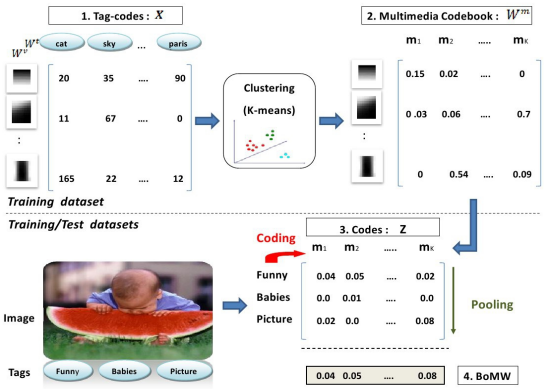

Figure 1. BoMW signature generation.

3.1 Coding

Different methods have been investigated in the literature in order to map local features to codes over the visual codebook, preserving some interesting properties such as sparsity [19], locality in the feature space [20], saliency [5], etc. These coding schemes alleviate the main drawbacks of classic coding ones namely hard and soft assignments [13, 15]. The locality based coding is currently the most interesting technique in terms of

trade-off between robustness and computational complexities. In [8] for instance, authors propose an efficient implementation of the locality-constrained coding of [20] by restricting the probabilistic soft coding of [15] approach to the k-nearest-codewords to a descriptor in the feature space. In our case, a tag-code $\mathbf{x}_i$ (a column of $\mathbf{X}$) of a given image is the descriptor to be coded over the M-codebook $\mathcal{W}^m$ as the following:

$$z_{i,j} = \begin{cases} \frac{\exp\left(-\beta||\mathbf{x}_i - \mathbf{m}_j||_2^2\right)}{\sum_{r=1}^{k} \exp\left(-\beta||\mathbf{x}_i - \mathbf{m}_r||_2^2\right)} & \text{if } \mathbf{m}_j \in \mathcal{N}_k(\mathbf{x}_i)\,, \\ 0 & \text{otherwise,} \end{cases} \tag{3}$$

where $\mathbf{z}_i$, a vector of size K^m, is the obtained code associated to the tag-code $\mathbf{x}_i$, $\mathcal{N}_k(\mathbf{x}_i)$ denotes the set of k-nearest neighbors to the vector $\mathbf{x}_i$ within the tag-code set of column vectors in $\mathbf{X}$ and β is a parameter controlling the weight decay speed of the locality.

3.2 Pooling

Given the coding coefficients of all tags within one image, a pooling operation is performed to obtain a compact signature $\mathbf{h}$, while preserving important information and discarding irrelevant details. This operation is formulated as the following:

$$h_j = g\Big(\left\{z_{i,j}\,; i = 1, ..., N\right\}\Big)\,; \forall j = 1, ..., K^m\,, \tag{4}$$

with g a pooling function such as the average, the sum or the maximum functions and N the number of tags associated to the current image caption. Recent works [2, 8] show, theoretically and empirically, that max-pooling is best suited to the recognition task. It is performed by selecting the maximum coding coefficient (or the salient codeword response) over tag-codes for each multimedia word.

4. Experiments

In this section, we report classification results on Pascal VOC 2007 [3], which consists of $9,963$ images (5011 for training and the rest for testing) annotated according to 20 classes. This benchmark is one of the most challenging because of the large variation on view size, illumination, scale, deformation and clutter, as well as complex backgrounds. Classification performances of the proposed system on the current dataset are evaluated in terms of (1) mean average precision (mAP), (2) computation cost for both signature design and classifier training and test, (3) stability of results toward codebook size. These three issues are the most challenging ones in the classification task, since they involve robustness of the recognition system and its scalability to large scale datasets.

4.1 Pipeline

We considered the following setup. Dense SIFT are extracted within a regular spatial grid at only one scale from images. The step-size is fixed to 6 pixels and patch size to 16×16 pixels. Visual codebooks of various sizes have been generated using the K-means clustering method on randomly selected SIFTs from the training set. A textual codebook of size 804 is generated using the same experiment setting as [4]. Once the tag-coding matrix has been created, the M-codebook is generated by clustering columns of the tag-codes matrix using K-means. We also fixed different sizes in order to analyze the robustness of the BoMW signature toward codebook size. When designing BoMW, coding tag-codes over the M-codebook is performed using the locality-constrained soft assignment with a neighborhood of size 5 and a softness parameter $\beta = 10$, similarly to [8]. Finally, we used a linear SVM for classification.

4.2 Discussion

Figure 2 shows classification performances in terms of mean average precision (mAP) using either BoVW or BoMW, while changing the sizes of visual and multimedia codebooks. For the BoVW, the spatial pyramid matching [7] is performed with three levels.

At any codebook size, classification results using BoMW outperform by about $6\% \sim 10\%$ those obtained with BoVW (Figure 2) as well as [4] that reported 43.3% with BoTW. This is expected since the proposed multimedia words are semantically higher than the low-level visual local features and the tag words separately, and also more consistent to encode the content of a multimedia document through an effective fusion of visual and text modes.

In opposition to classic BoW signatures, classification results remain stable aaccording to the size of the visual/multimedia codebooks. This behavior, that reduces the complexity of the classification system both during training and testing, is obtained at the cost of a small pre-processing step for signature design (building tag-coding matrix and clustering it). The best classification scores obtained with the BoVW and BoTW are 49.36% and 43.3% respectively, using visual signatures of size 86016 and textual signatures of size 804. The best classification score obtained with the proposed BoMW is 55.54% using a signature of size 512 (see Figure 2). The performance gain is due to the fact that the proposed multimedia signatures lie on a structured space, well appropriate to describe multimedia documents. Therefore, BoMW seem more class-discriminative than other types of BoW.

We argue that the proposed multimedia signature carries interesting supplementaty information in comparison to the "monomedia" features. To address this point, we combined the three types of BoW signatures (textual, visual, multimedia) through a late fusion scheme. On the Pascal VOC 2007 base, this method resulted into a mAP of $\mathbf{67.78}\%$ that outperform the best score reported on this benchmerk (66.7% in [4]). Moreover, let notice that the proposed method brings a significant reduction of both memory use and computation complexity. [4] fused one textual signature and 15 local and global visual signatures while we used only three compact BoW signatures (BoVW, BoTW and BoMW of sizes 86016, 804 nd 512 respectively). As well, [4] used *Multiple Kernel Learning* to learn theconcept while ours result from a simpler linear kernel combined with *stacked generalization* [18].

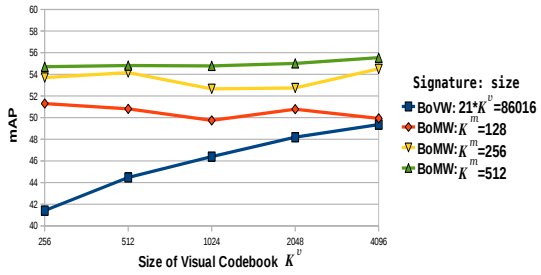

Figure 2. Results on Pascal VOC 2007.

5. Conclusion

We introduced a new BoW based signature that is appropriate to describe multi-modal documents. The design of the proposed signature takes advantage of the recent advances in generating Bag-of-Visual-Words. Experiments have been conducted on a well-known challenging benchmark. They show the competitive performances of the Bag-of-Multimedia-Words, ensuring a trade-off between classification accuracy and computation cost. The proposed framework is generic and, in the future, we plan to exploit it in other application domains (video classification, robotics etc.), with data that include other modes than textual and visual ones.

6 Acknowledgments

A. Znaidia is supported by grants from DIGITEO and Région Ile-de-France. This work has been partially funded by I2S in the context of the project Polinum.

References

[1] S. Battiato, G. M. Farinella, G. Gallo, and D. Ravì. Exploiting textons distributions on spatial hierarchy for scene classification. *J. Image Video Process.*, 2010.
[2] Y.-L. Boureau, F. Bach, Y. LeCun, and J. Ponce. Learning mid-level features for recognition. In *CVPR*, pages 2559–2566, 2010.
[3] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. 'the pascal visual object classes challenge 2007 (voc2007) results.
[4] M. Guillaumin, J. Verbeek, and C. Schmid. Multimodal semi-supervised learning for image classification. In *CVPR*, pages 902 – 909, 2010.
[5] Y. Huang, K. Huang, Y. Yu, and T. Tan. Salient coding for image classification. In *CVPR*, 2011.
[6] L. S. Kennedy, S. fu Chang, and I. V. Kozintsev. To search or to label?: predicting the performance of search-based automatic image classifiers. In *MIR*, pages 249–258, 2006.
[7] S. Lazebnik, C. Schmid, and J. Ponce. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In *CVPR*, pages 2169–2178, 2006.
[8] L. Liu, L. Wang, and X. Liu. In Defense of Soft-assignment Coding. In *ICCV*, 2011.
[9] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.
[10] F. Monay and D. Gatica-Perez. pLSA-based image auto-annotation: constraining the latent space. In *ACM Multimedia*, pages 348–351, 2004.
[11] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, New York, 1983.
[12] B. Sigurbjörnsson and R. van Zwol. Flickr tag recommendation based on collective knowledge. In *WWW*, pages 327–336, 2008.
[13] J. Sivic and A. Zisserman. Video Google: A Text Retrieval Approach to Object Matching in Videos. In *ICCV*, volume 2, pages 1470–1477, 2003.
[14] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *PAMI*, 22:1349–1380, 2000.
[15] J. van Gemert, C. Veenman, A. Smeulders, and J. Geusebroek. Visual word ambiguity. *PAMI*, 2009.
[16] G. Wang, D. Hoiem, and D. Forsyth. Building text features for object image classification. In *CVPR*, 2009.
[17] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In *CVPR*, 2010.
[18] D. H. Wolpert. Stacked generalization. *Neural Networks*, 5:241–259, 1992.
[19] J. Yang, K. Yu, Y. Gong, and T. Huang. Linear spatial pyramid matching using sparse coding for image classification. In *CVPR*, 2009.
[20] K. Yu, T. Zhang, and Y. Gong. Nonlinear learning using local coordinate coding. *NIPS*, 2009.
[21] W.-L. Zhao, Y.-G. Jiang, and C.-W. Ngo. Keyframe Retrieval by Keypoints: Can Point-to-Point Matching Help? In *CIVR*, 2006.