Sampling of Web Images with Dictionary Coherence for Cross-Domain Concept Detection

Yongqing Sun, Kyoko Sudo, Yukinobu Taniguchi, and Masashi Morimoto

NTT Media Intelligence Laboratories,
1-1 Hikarinooka Yokosuka-shi Kanagawa, 239-0847, Japan
yongqing.sun@lab.ntt.co.jp

Abstract. Due to the existence of cross-domain incoherence resulting from the mismatch of data distributions, how to select sufficient positive training samples from scattered and diffused web resources is a challenging problem in the training of effective concept detectors. In this paper, we propose a novel sampling approach to select coherent positive samples from web images for further concept learning based on the degree of image coherence with a given concept. We propose to measure the coherence in terms of how dictionary atoms are shared since shared atoms represent common features with regard to a given concept and are robust to occlusion and corruption. Thus, two kinds of dictionaries are learned through online dictionary learning methods: one is the concept dictionary learned from key-point features of all the positive training samples while the other is the image dictionary learned from those of web images. Intuitively, the coherence degree is then calculated by the Frobenius norm of the product matrix of the two dictionaries. Experimental results show that the proposed approach can achieve constant overall improvement despite cross-domain incoherence.

Keywords: Visual concept detection, Semantic indexing, Web image mining, Sparse representation, Dictionary learning.

1 Introduction

Nowadays, the explosive growth of visual contents on the Internet presents a challenge in how to manage the ever-growing size of the multimedia collections, particularly in how to extract sufficiently accurate semantic metadata (concepts) to make them searchable [5]. Visual concept detection is essentially a classification task in which classifiers are learned with various features extracted from training samples to predict the presence of a certain concept in a video shot or keyframe (image) [14, 15]. Ranging from objects such as *"boat"* and *"car"* to scenes such as *"sky"* and *"sea"*, semantic concepts can serve as good intermediate semantic metadata for video content indexing and understanding [14]. With a large set of robust concept detectors, significant improvement can be achieved in many challenging applications, such as image/video search and summarization [15].

S. Li et al. (Eds.): MMM 2013, Part II, LNCS 7733, pp. 283–292, 2013.

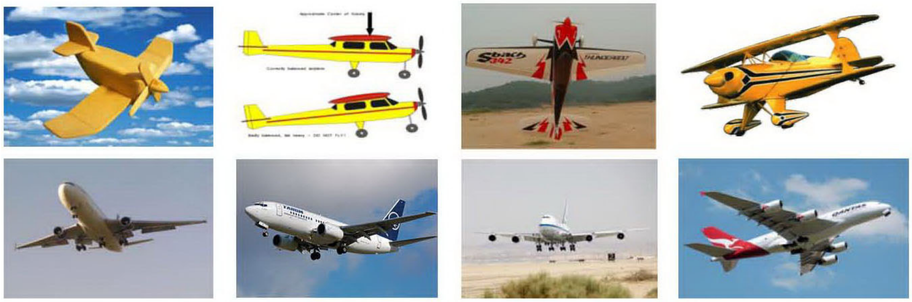

Fig. 1. Web Image Examples of "Airplane-flying"

In order to learn effective concept detectors, a critical step is to acquire a sufficiently large amount of training samples, especially positive training samples [5]. Fortunately, with the explosive growth of visual contents on the Internet, large amounts of training samples have become available through Web searching [2, 15]. Consequently, how to utilize these abundant web images to improve concept detection has been the subject of intensive research by a large multimedia research community, since it has offered promising ways to automatically annotate the contents at relatively low cost [2, 15]. [15] empirically studied the effect of exploiting tagged images on concept learning by analyzing tag lists. [2] proposed an automatic concept-to-query mapping method for acquiring training data from online platforms.

However, the online web images are very noisy, cover a wide range of unpredictable contents, and have quite different data distributions with any close dataset such as TREC-Vid dataset [9, 13]. As shown in Figure 1, for example, the content of web images searched from Google Image with the keyword "Airplane-flying" varies greatly. Obviously, the images in the top row of the figure are incoherent from the concept "Airplane-flying" in the TRECVid dataset. Thus these images can not facilitate the training of the concept and may even harm it. Only the images in the bottom row are consistent with the dataset and hence helpful. Therefore, how to select coherent positive training samples from diffused web images is a challenging problem for training of effective concept detectors [2, 11, 12] due to the existence of cross-domain incoherence resulting from the mismatch of data distributions.

Existing work on video concept learning using web images has mainly focused on how to leverage compact features, such as region-based features [12] or image salience [11], to alleviate the visual differences. Since an image is greatly reduced to a very compact feature vector, the effect of these approaches is not evident. In this paper, we propose a novel sampling approach on how to exploit bundles of local key-point features to measure how coherent a web image is with a given concept, from the aspects of sparse coding and dictionary learning.

2 Sparse Coding and Dictionary Learning

Recently, modeling data or signals as sparse linear combinations of a few elements (atoms) of some redundant bases (dictionary), sparse coding or sparse representation has been widely applied to classification problems where the data on multiple subspaces relies on the notion of sparsity due to its robustness to occlusion and corruption [8].

Formally, given a dictionary $\mathbf{D} = [d_1, \ldots, d_k] \in \mathbf{R}^{l \times k}$, and the i-th data instance vector $x_i \in \mathbf{R}^l$ from the observed data matrix $\mathbf{X} = [x_1, \ldots, x_n] \in \mathbf{R}^{l \times n}$, where d_i is the dictionary atom, l is the data dimensionality, k is the size of the dictionary, and n is the number of data instances ($l < k \ll n$), sparse representation solves the following non-convex program to seek the sparsest solution for the coefficient vector (i.e., sparse code) $\alpha_i \in \mathbf{R}^k$:

$$\min_{\alpha_i} \| \alpha_i \|_0 \;\; s.t. \, \mathbf{D}\alpha_i = x_i \tag{1}$$

where $\| \alpha_i \|_0$ denotes the l_0 pseudo-norm of the coefficient vector $\alpha_i \in \mathbf{R}^k$, i.e., the number of non-zero elements.

Since minimizing l_0 is NP-hard, a common approximation is to replace it with the l_1-norm according to theories from compressive sensing [3]. Taking noise into consideration, the equality constraint must be relaxed. Hence, an alternative is to solve the unconstrained problem after using the Lagrange multiplier method:

$$\min_{\alpha_i} \frac{1}{2} \parallel x_i - \mathbf{D}\alpha_i \parallel^2 + \lambda \parallel \alpha_i \parallel_1, \tag{2}$$

where λ is a regularization parameter that balances the tradeoff between reconstruction error and sparsity induced by the alternative l_1-norm constraint. This is a convex problem called Lasso in statistics and can be efficiently solved by the LARS-Lasso algorithm [4].

Here, how to determine the dictionary $\mathbf{D}$ is very important for sparse representation. It has been shown that dictionaries learned from data can significantly outperform off-the-shelf ones such as wavelets [8]. Given the observed data matrix $\mathbf{X}$, the goal is to seek an optimal $\mathbf{D}$ so that all the data instances can be represented as a sparse linear combination of their atoms. There are many dictionary learning methods such as the method of optimal directions (MOD), the K-SVD algorithm , and the Generalized Principal Component Analysis (GPCA) [8]. All the methods using classical optimization alternate between the dictionary and sparse code, and can obtain good results, but are too slow to scale up to large data sets [8]. Recently, efficient online learning methods were proposed in [8], which can handle large scale, potentially infinite, or dynamic data sets.

3 Proposed Approach

3.1 Overview

Inspired by the observation that dictionary atoms representing common features in all categories tend to appear to be repeated almost exactly in dictionaries

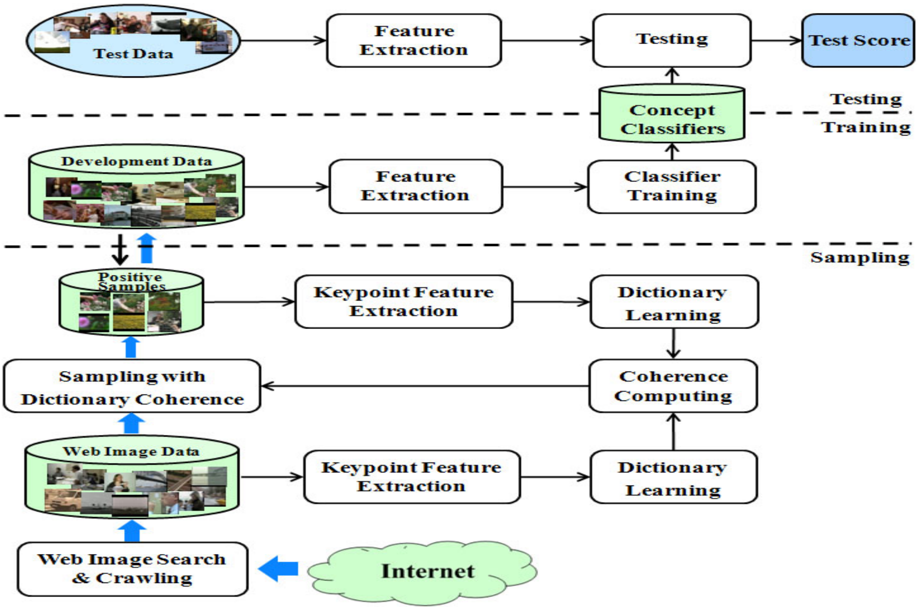

Fig. 2. Proposed Framework

corresponding to different categories, [10] promotes incoherence between the dictionary atoms to improve the speed and accuracy of sparse coding.

Motivated by this work, since the shared dictionary atoms learned from data can represent common features with regard to a given concept (represented by the set of positive training samples) and are robust to occlusion and corruption [8], we propose to use dictionary coherence in terms of how an image and a given concept share dictionary atoms to measure the degree of image coherence with the concept. That is, the more atoms they share, the higher the dictionary coherence is, which means it is more probable that the web image is coherent with the concept.

In order to compute the dictionary coherence, we learn two kinds of dictionaries through the online dictionary learning method [8]: one is the concept dictionary learned from key-point features of all the positive training samples while the other is the image dictionary learned from those of web images. Intuitively, the coherence degree is then calculated by the Frobenius norm of the product matrix of the two dictionaries since it reflects the sum of the absolute values of inner products between dictionary atoms.

On the basis of the dictionary coherence, we propose a novel adaptive sampling approach to select coherent positive samples from diffused web images for further concept learning.

3.2 Algorithm

As shown in the framework of Figure 2, for each concept, the algorithm of the proposed sampling principally consists of the following steps:

(1) Construction of concept set: Select all the positive training samples from a development dataset such as TRECVid development set to represent the concept.
(2) Feature extraction of concept set: Extract local key-point features, such as SIFT [7] or SURF [1], and collect each key-point feature $x_i \in \mathbf{R}^l$ of all the images in the concept set to form the data matrix $\mathbf{X}_C = [x_1, \ldots, x_n] \in \mathbf{R}^{l \times n}$. Here, l is the feature dimensionality, and n is the total number of keypoints.
(3) Concept dictionary learning: Adopt the efficient online dictionary learning methods [8] to learn the concept dictionary $\mathbf{D}_C \in \mathbf{R}^{l \times k}$ from the concept data matrix $\mathbf{X}_C$, where k is the size of the dictionary, i.e., the number of atoms. For the SIFT feature, we set $k = 192$ about 1.5 to 2.0 times of the feature size $l = 128$ [14].
(4) Collection of web image set: After query construction or mapping [2] based on the concept name, search the web images and crawl the top-ranked ones.
(5) Feature extraction of web image: For each image in the web image set, extract the same local key-point features as the second step, and form the image data matrix $\mathbf{X}_{\dot{\imath}} \in \mathbf{R}^{l \times m}$, where m is the number of keypoints in the image.
(6) Image dictionary learning: Adopt the same dictionary learning methods [8] to learn the image dictionary $\mathbf{D}_{\dot{\imath}} \in \mathbf{R}^{l \times k}$ from the image data matrix $\mathbf{X}_{\dot{\imath}}$.
(7) Dictionary coherence computing: Use Equation (4) in subsection 3.4 to compute the dictionary coherence C_i between the image dictionary $\mathbf{D}_{\dot{\imath}}$ and the concept dictionary $\mathbf{D}_C$.
(8) Adaptive sampling: Compare the dictionary coherence C_i of the current web image with the adaptive threshold in subsection 3.5 to determine whether to add the current web image to the training set.

As shown in Figure 2, after adding the selected coherent positive web samples (a manual check is advised to ensure it is positive) to the training set, we can do further concept learning for training more effective concept detectors. We will detail the key procedures in the following subsections.

3.3 Dictionary Learning

In our study, we use the efficient online learning methods [8] to learn the dictionary. Due to the advantage of non-negativity constraints in learning part-based representations [14], which is helpful for object-oriented concept learning, we impose the positivity constraints on both dictionary D and sparse code α_i in solving the optimization problem as below:

$$\min_{\mathbf{D},\alpha_i} \sum_{i=1}^{n} \left(\frac{1}{2} \parallel x_i - \mathbf{D}\alpha_i \parallel^2 + \lambda \parallel \alpha_i \parallel_1 \right), \; s.t., \mathbf{D} \geq 0, \, \alpha_i \geq 0. \tag{3}$$

while restricting the atoms to have a norm of less than one. The optimization is achieved through an iterative approach consisting of two alternative steps: the sparse coding step on a fixed $\mathbf{D}$ and the dictionary update step on fixed α_i [8]. As mentioned above, we learn two types of dictionaries: (1) a concept dictionary $\mathbf{D}_C$; (2) an image dictionary $\mathbf{D}_i$.

3.4 Dictionary Coherence Computing

The natural way to measure the degree of coherence C_i between the image dictionary $\mathbf{D}_i$ and the concept dictionary $\mathbf{D}_C$, is to inspect the product matrix: $\mathbf{D}_i^{\mathbf{T}}\mathbf{D}_C$, where the superscript T denotes the matrix transposition. This is because the element d_{ij} of the product matrix represents the inner product between a pair of the two dictionary atoms, i.e., $d_{ij} = d_i \cdot d_j$, here, $d_i \in \mathbf{D}_i$, $d_j \in \mathbf{D}_C$. Therefore, as shown in Equation (4), we compute dictionary coherence C_i through a Frobenius norm defined as the square root of the sum of the absolute squares of the matrix's elements d_{ij}:

$$C_i = \| \mathbf{D}_i^{\mathbf{T}}\mathbf{D}_C \|_F = \sqrt{\sum_{i=1}^{k}\sum_{j=1}^{k} |d_{ij}|^2} \tag{4}$$

where the subscript F denotes the Frobenius norm.

3.5 Adaptive Sampling

After computing the dictionary coherence C_i between the current web image and the concept, we can easily determine whether to add the current web image to the training set by simply comparing the C_i with a pre-given threshold C_{th}. If $C_i \geq C_{th}$, meaning that the web image is coherent with the concept, then we accept it. Otherwise, we discard it.

Here, we propose an adaptive off-line method through automatic calculation of the threshold C_{th} from the distribution of the coherence degrees of all the positive train samples. According to the theory of hypothesis testing, the threshold C_{th} can be adaptively determined by:

$$C_{th} = \mu - \eta\sigma, \tag{5}$$

where μ and σ are the mean and standard deviation of all the coherence degrees C_{Pos} between each positive training sample and the concept, and η is an empirical parameter that can be determined universally. In our experiments, we set $\eta = \sqrt{3}$.

4 Experimental Results

We tested the proposed method on the TRECVid 2008 [9, 13]. TRECVid is now widely regarded as the actual standard for performance evaluation of concept based video retrieval systems [14]. The number of positive training samples

Table 1. The number of positive samples for 20 concepts in TRECVid 08

ID	Concept	#DPos	#WPos	#SPos
1001	Classroom	241	790	347
1002	Bridge	186	420	235
1003	Emergency-Vehicle	103	151	11
1004	Dog	136	795	123
1005	Kitchen	289	537	174
1006	Airplane-flying	80	395	113
1007	Two-people	4140	729	458
1008	Bus	106	902	312
1009	Driver	302	489	157
1010	Cityscape	331	879	623
1011	Harbor	217	261	76
1012	Telephone	203	557	412
1013	Street	1799	693	508
1014	Demonstration-Or-Protest	159	68	25
1015	Hand	1879	384	302
1016	Mountain	265	507	284
1017	Nighttime	490	594	229
1018	Boat-Ship	506	783	215
1019	Flower	620	948	513
1020	Singing	441	646	187

Note: The column “#DPos” denotes the number of positive training samples in the TRECVid 08 development set, “#WPos” in the initial positive web image set, “#WPos” in the final web image set after sampling.

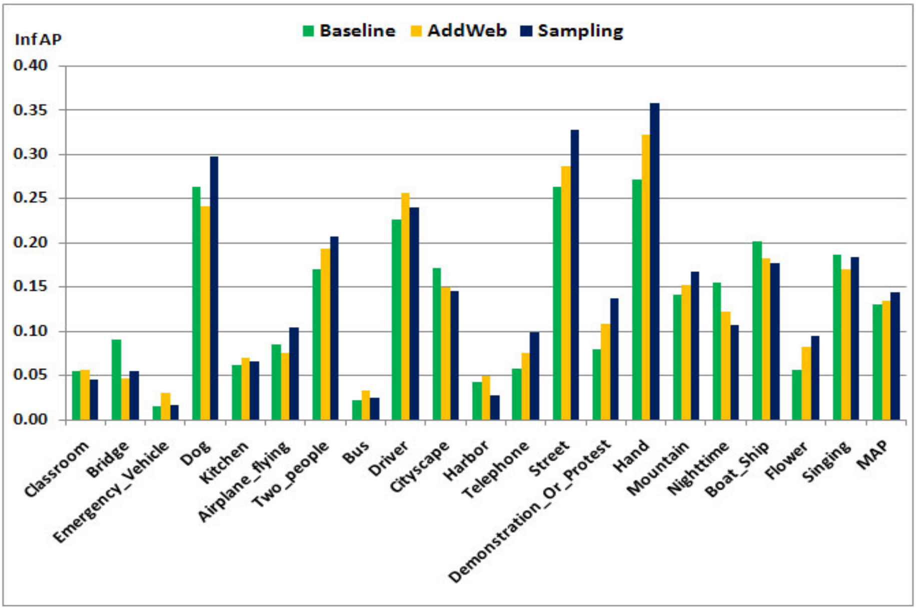

Fig. 3. Comparison results

for each concept in the TRECVid 08 development set is shown in the column "#DPos" of Table 1 [14].

First, we used the Google API to search and download the top 1000 web images for each concept by constructing a query with the concept name. Then we annotated the images manually; the number of positive samples for each concept in the initial web image set is shown in the column "#WPos" of Table 1. Finally, we used our proposed sampling method to select the positive samples for each concept; the number of positive samples for each concept selected from the web images is shown in the column "#SPos" of Table 1. To test the effectiveness of our proposed method, we performed three runs for each concept:

- [**Baseline**]: Use only positive training samples in the TREC-Vid 08 development set ("#DPos" in Table 1).
- [**AddWeb**]: Use positive training samples of the TREC-Vid 08 development set and the initial positive web image set ("#DPos+#WPos" in Table 1).
- [**Sampling**]: Use positive training samples of the TREC-Vid 08 development set and the web image set after the proposed sampling("#DPos+#SPos" in Table 1).

In the above runs, we used the SIFT features [7] for dictionary learning during sampling, and the well-known BoW feature [6] based on soft-weighting of SIFT, due to its widely reported effectiveness [14].

Figure 3 shows the comparison results of AP for each concept and mean AP (MAP) of the three runs. As shown, the proposed run [Sampling] achieved the highest MAP of 0.144, which is 9.92% higher than the run [Baseline] (MAP 0.131), and 6.67% higher than the run [AddWeb] without sampling(MAP 0.135). In particular, the proposed method outperformed the others on 9 out of 20 concepts, including Airplane-flying, Dog, Telephone, Demonstration-Or-Protest, Hand, and Flower, which had been selected with sufficient visually-coherent positive samples, while little was gained with the concepts such as Harbor, Kitchen, Bridge, and Emergency-Vehicle because these concepts on the old documentary TRECVid videos may be too outdated for enough positive web samples to be obtained. On the other hand, the run [AddWeb] achieved only a 3.05% improvement in MAP compared with the run [Baseline].

Compared with the best runs in TRECVid 2008 [6], significant improvement was obtained in handling concepts with few TRECVID positive training samples. The experimental results show that the proposed approach can achieve constant overall improvement despite cross-domain incoherence.

5 Conclusion

In this paper, we proposed a novel sampling of web images for cross-domain concept detection based on coherence between an image dictionary and concept dictionary. Experimental results on the TRECVid 08 benchmark show the effectiveness and necessity of the proposed sampling method.

References

1. Bay, H., Ess, A., Tuytelaars, T., Van Gool, L.: Surf: Speeded up robust features. Computer Vision and Image Understanding 110(3), 346–359 (2008)
2. Borth, D., Ulges, A., Breuel, T.M.: Automatic concept-to-query mapping for web-based concept detector training. In: ACM Multimedia 2011, New York, USA, pp. 1453–1456 (2011)
3. Donoho, D.: For most large underdetermined systems of linear equations the minimal l1-norm solution is also the sparsest solution. Comm. Pure and Applied Math. 59(6), 797–826 (2006)
4. Efron, B., Hastie, T., Johnstone, I., Tibshirani, R.: Least angle regression. Annals of Statistics 32(2), 407–499 (2004)
5. Huiskes, M.J., Thomee, B., Lew, M.S.: New trends and ideas in visual concept detection: the mir flickr retrieval evaluation initiative. In: MIR 2010, New York, USA, pp. 527–536 (2010)
6. Jiang, Y.-G., Yang, J., Ngo, C.-W., Hauptmann, A.G.: Representations of keypoint-based semantic concept detection: A comprehensive study. IEEE Transactions on Multimedia 12, 42–53 (2010)
7. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)
8. Mairal, J., Bach, F., Ponce, J., Sapiro, G.: Online learning for matrix factorization and sparse coding. J. Mach. Learn. Res. 11, 19–60 (2010)
9. Over, P., Awad, G., Rose, R.T., Fiscus, J.G., Kraaij, W., Smeaton, A.F.: Trecvid 2008 - goals, tasks, data, evaluation mechanisms and metrics. In: TRECVID Workshop (2008)

10. Ramirez, I., Sprechmann, P., Sapiro, G.: Classification and Clustering via Dictionary Learning with Structured Incoherence and Shared Features. In: CVPR 2010, pp. 3501–3508 (June 2010)
11. Sun, Y., Kojima, A.: A novel method for semantic video concept learning using web images. In: ACM Multimedia 2011, New York, USA, pp. 1081–1084 (2011)
12. Sun, Y., Shimada, S., Taniguchi, Y., Kojima, A.: A novel region-based approach to visual concept modeling using web images. In: ACM Multimedia 2008, New York, USA, pp. 635–638 (2008)
13. Tang, S., Li, J.-T., Li, M., Xie, C., Liu, Y.-Z., Tao, K., Xu, S.-X.: TRECVID 2008 High- Level Feature Extraction By MCG-ICT-CAS. In: Proc. TRECVID 2008 Workshop, Gaithesburg, USA (November 2008)
14. Tang, S., Zheng, Y.-T., Wang, Y., Chua, T.-S.: Sparse ensemble learning for concept detection. IEEE Transactions on Multimedia 14(1) (2012)
15. Zhu, S., Wang, G., Ngo, C.-W., Jiang, Y.-G.: On the sampling of web images for learning visual concept classifiers. In: CIVR 2010, New York, USA, pp. 50–57 (2010)