ISI at ImageCLEF 2012: Scalable System for Image Annotation

Yoshitaka Ushiku, Hiroshi Muraoka, Sho Inaba, Teppei Fujisawa, Koki Yasumoto, Naoyuki Gunji, Takayuki Higuchi, Yuko Hara, Tatsuya Harada, and Yasuo Kuniyoshi

Intelligent Systems and Informatics Lab., the University of Tokyo
{ushiku,muraoka,inaba,fujisawa,
k-yasumoto,gunji,higuchi,y-hara,
harada,kuniyosh}@isi.imi.i.u-tokyo.ac.jp
http://www.isi.imi.i.u-tokyo.ac.jp

Abstract. We participate in the ImageCLEF 2012 Photo Annotation Tasks. We devote our attention to make our system scalable for the data amount. Therefore we train linear classifiers with our online multilabel learning. For Flickr Photo task, we extract visual Fisher Vectors (FVs) from some kinds of local descriptors and used the provided Flickr-tags for textual features. For Web Photo tasks, we just use the provided Bag-of-Visual-Words (BoVW) of some kinds of SIFT descriptors. A linear classifier for each label is obtained with an online multilabel learning, Passive-Aggressive with Averaged Pairwise Loss (PAAPL). The results have shown that our scalable system achieves pretty good performances in all tasks we take part in.

1 Introduction

In this paper, we describe our method for the ImageCLEF 2012 Photo Annotation tasks. In particular, we attack three tasks: concept annotation using Flickr photos [8], improving performance in Flickr concept annotation task using Web photos [10], and scalable concept image annotation using Web photos [10].

Especially, we pay our attention to the scalability for the data amount. In this literature, many techniques are developed to improve the performance of object recognition. Though some of them have succeeded by introducing a complicated classifier such as the multiple kernel SVM, the complexity for leaning and annotating is a problem. Because many kinds of labels require a large amount of training data, the scalability for the data amount is important for generic object recognition.

Consequently, our objective is to investigate scalable methods for feature extraction, for learning, and for annotation. Recent studies for large scale image classification adopt online learning for linear classification. In [7, 4], high-dimensional features in [6, 11] are used for learning 1000 classes from over a million images. In fact, Fisher Vectors (FVs) and linear SVM won the ImageCLEF 2010 as described in [5].

Our main contribution is the investigation of our novel online learning for multilabel problem. Because batch learning by loading all training samples is impossible, usage of an online learning is a promising method in order to realize scalability. In [7, 4], online SVM learning with Stochastic Gradient Descent [1] (SGD-SVM) is applied with a one-vs.-the-rest manner. The classifier for a label is obtained by regarding images associated with the label as positive samples and the rest images as negative samples. Furthermore, labels are output according to the scores from the binary classifiers. Nevertheless, no guarantee exists that the output of SVMs for different classifiers will have appropriate scales. Thus we investigate a multiclass learning Passive-Aggressive algorithm [2] to solve this problem. In [9], we have proposed Passive-Aggressive with Averaged Pairwise Loss (PAAPL) for which multiple labels are attached to one sample. At first, we use an averaged pairwise loss instead of the hinge-loss of PA. Secondly, we randomly select these pairs at every learning. These two improvements make PAAPL can converge faster than PA.

2 Feature Extraction

In this section, we describe the features we use in three tasks. For the Flickr Photo task, we extract FVs as visual features and some kinds of BoW of Flickr-tags as textual features. For the Web Photo tasks, we use the provided Bag-of-Visual-Words only.

2.1 Visual Features

Bag-of-Visual-Words. Bag-of-Visual-Words (BoVW) is quite a popular approach for image classification, because it achieves good performance in spite of its simplicity. The main idea is that images are treated as loose collections of K codewords, representative local descriptors, and that each key-point patch, in which a local descriptor is extracted, is sampled independently. The BoVW feature is obtained by making a histogram of the number of local descriptors assigned to each codeword. The dictionary, which consists of K codewords, has to be generated by unsupervised clustering of training samples in advance and each local descriptor is assigned to the nearest codeword in the dictionary. BoVW vector is therefore K-dimensional.

Fisher Vectors. Fisher Vectors (FV), which is regarded as an extension of the BoVW representation, is a standard approach to the large-scale image recognition. BoVW utilizes 0-order statistics of the distribution of local descriptors, whereas FV utilizes 1- and 2-order statistics. The distribution of local descriptors is fitted to the mixture model of K Gaussians, and the gradients of the likelihood in the parameter space are computed. The gradients describe which direction the model parameters are to be modified to get a better description of the image. The dimensions are then whitened by multiplying the square root of the Fisher information matrix.

We denote T local descriptors by $X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_T\}$, and the mixture weight, mean, covariance matrix of i-th Gaussian by w_i, μ_i, and σ_i, respectively. Since the covariance matrices are assumed to be diagonal, we denote the variance vector by σ^2. The FV representation is thus given as,

$$\mathcal{G}^X_{\mu,i} = \frac{1}{T\sqrt{w_i}} \sum_{t=1}^{T} \gamma_t(i) \left(\frac{x_t - \mu_i}{\sigma_i} \right), \tag{1}$$

$$\mathcal{G}^X_{\sigma,i} = \frac{1}{T\sqrt{2w_i}} \sum_{t=1}^{T} \gamma_t(i) \left[\frac{(x_t - \mu_i)^2}{\sigma_i^2} - 1 \right], \tag{2}$$

where $\gamma_t(i)$ is the soft assignment of $\boldsymbol{x}_t$ to i-th Gaussian as follows,

$$\gamma_t(i) = \frac{w_i u_i(x_t)}{\Sigma_{j=1}^{K} w_j u_j(x_t)}. \tag{3}$$

Then, we obtain the 2KD-Dimensional vector by concatenating $\mathcal{G}^X_{\mu,i}$ and $\mathcal{G}^X_{\sigma,i}$. To enhance the performance, a power normalization is proven to be effective in [6]. The vectors are normalized with L_2 norms after the power normalization.

2.2 Text Features

We use Bag-of-Words (BoW) representation, which is based on the idea that each word in a text appears independently. BoW is obtained by counting appearance of words in a text. In our method, the feature is converted in following two ways, TF-IDF and L_2-normalization.

TF-IDF weight We regard the typicality of each Flickr-tag as a clue to how the tag relates to the image's contents. Therefore, we use TF-IDF value for each element of a BoW vector.

***L*$_2$-normalization** To reduce the effect of different numbers of tags among images, we simply L_2-normalize the BoW vectors.

3 Online Multilabel Learning

To learn the models for each label from various images, requirements are not only compatibility of scalability for the data amount and accuracy for label estimation, but also tolerability of noise.

Given the t-th training sample $\boldsymbol{x}_t \in \mathbb{R}^d$ associated with a label set Y_t, a subset of $\mathcal{Y} = \{1, \ldots, n_y\}$, it is classified with the present weight vector $\boldsymbol{\mu}_t^{y_i}$ $(i = 1, \ldots, n_y)$[1] as,

$$\hat{y}_t = \arg\max_{y_i} \boldsymbol{\mu}_t^{y_i} \cdot \boldsymbol{x}_t. \tag{4}$$

[1] Here, the bias b is included in $\boldsymbol{\mu}_t$ as $\boldsymbol{\mu}_t^\top \leftarrow [\boldsymbol{\mu}_t^\top, b]$ by redefining $\boldsymbol{x}_t^\top \leftarrow [\boldsymbol{x}_t^\top, 1]$

If necessary, multiple labels are estimated in score order.

Multilabeling for one sample is applicable by defining $n_y > 1$. Here, hinge-loss ℓ is given as,

$$\ell(\boldsymbol{\mu}_t^{r_t}, \boldsymbol{\mu}_t^{s_t}; (\boldsymbol{x}_t, Y_t)) = \begin{cases} 0 & \boldsymbol{\mu}_t^{r_t} \cdot \boldsymbol{x}_t - \boldsymbol{\mu}_t^{s_t} \cdot \boldsymbol{x}_t \geq 1 \\ 1 - (\boldsymbol{\mu}_t^{r_t} \cdot \boldsymbol{x}_t - \boldsymbol{\mu}_t^{s_t} \cdot \boldsymbol{x}_t) & \text{otherwise} \end{cases}. \quad (5)$$

where $r_t = \arg\min_{r \in Y_t} \boldsymbol{\mu}_t^r \cdot \boldsymbol{x}_t$ and $s_t = \arg\max_{s \notin Y_t} \boldsymbol{\mu}_t^s \cdot \boldsymbol{x}_t$.

PA is an online learning method for binary and multiclass classification, regression, uniclass estimation and structure estimation. The biggest benefit of PA is that the update coefficient is analytically calculated according to the loss. In contrast, SGD based methods and traditional perceptron require designing the coefficient.

Here we seek to decrease the hinge-loss of multi-classification and not to change the weight radically. Consequently, we obtain the following formulation.

$$\boldsymbol{\mu}_{t+1}^{r_t}, \boldsymbol{\mu}_{t+1}^{s_t} = \arg\min_{\boldsymbol{\mu}^{r_t}, \boldsymbol{\mu}^{r_t}} ||\boldsymbol{\mu}^{r_t} - \boldsymbol{\mu}_t^{r_t}||^2 + ||\boldsymbol{\mu}^{s_t} - \boldsymbol{\mu}_t^{s_t}||^2 + C\xi^2, \quad (6)$$

$$\text{s.t. } \ell(\boldsymbol{\mu}^{r_t}, \boldsymbol{\mu}^{s_t}; (\boldsymbol{x}_t, Y_t)) \leq \xi \text{ and } \xi \geq 0. \quad (7)$$

Therein, ξ denotes a slack variable representing the bound of the loss. C signifies a parameter to reduce the negative influence of noisy labels. It can be derived using Lagrange's method of undetermined multipliers. Therefore we obtain,

$$\boldsymbol{\mu}_{t+1}^{r_t} = \boldsymbol{\mu}_t^{r_t} + \tau_t \cdot \boldsymbol{x}_t, \ \boldsymbol{\mu}_{t+1}^{s_t} = \boldsymbol{\mu}_t^{s_t} - \tau_t \cdot \boldsymbol{x}_t, \quad (8)$$

$$\tau_t = \min\{C, \ell(\boldsymbol{\mu}_t^{r_t}, \boldsymbol{\mu}_t^{s_t}; (\boldsymbol{x}_t, Y_t))/(2\boldsymbol{x}_t^2)\}. \quad (9)$$

This PA is called PA-II in [2]. PA and SGD-SVM have a closed form. Indeed, PA for binary classification and SGD-SVM without L_2 regularization have the same update rule. Differences between SGD-SVM and PA here are (1) binary or multi-class, (2) regularization form, and (3) the number of parameter to be tuned.

3.1 Passive-Aggressive with Averaged Pairwise Loss

PA is online learning methods for classification, but it presents no problem if a sample is associated with multiple labels. Indeed, the Passive–Aggressive Model for Image Retrieval (PAMIR) [3] is proposed by application of PA to image retrieval.

However, they treat only one relevant label and one irrelevant label. Apparently, models of some labels are not well updated and that convergence becomes delayed.

Therefore, we have proposed a novel online learning algorithm for which multiple labels are attached to one sample in [9]. General online learning methods consist of two steps: classification of the t-th sample, and update of the t-th models. Given the d-dimensional weight vectors $\boldsymbol{\mu}$ for all n_y labels, the complexity for

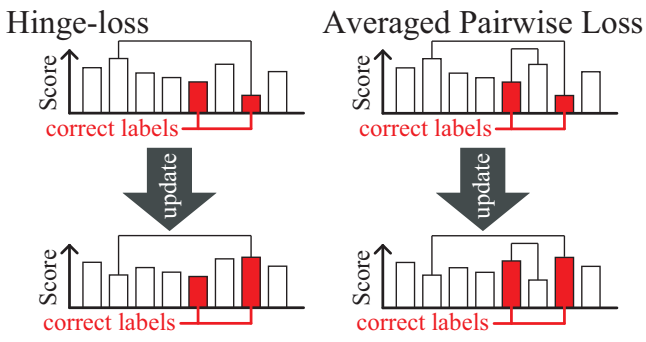

Fig. 1. Comparison between hinge-loss and averaged pairwise loss.

classification of a sample is $O(dn_y)$, while update of a model is $O(d)$. If we update all models with given labels Y_t, its complexity becomes $O(d|Y_t|)$. In image annotation and especially sentence generation, we can assume $n_y \gg |Y_t|$. Therefore, since classification is the rate-controlling step, total computation time remains much the same whether we update one model or $|Y_t|$ models. Fig.1 shows the conceptual difference between hinge-loss and the loss used in proposed method. Thus the proposed PAAPL achieves efficiency by averaging all pairwise loss between relevant and irrelevant labels.

1. Given a t-th image, define label set $\bar{Y}_t$ of n_y labels by selecting highly scored and irrelevant labels.
2. Randomly select one relevant label r_t from Y_t and one irrelevant label s_t from $\bar{Y}_t$.
3. Based on a hinge-loss between r_t and s_t, $1 - (\boldsymbol{\mu}_t^{r_t} \cdot \boldsymbol{x}_t - \boldsymbol{\mu}_t^{s_t} \cdot \boldsymbol{x}_t)$, update models according to PA.

Additionally, we investigate a way to reduce the complexity $O(dn_y)$ for the classification step. In [12], the approximation of a loss function by the random selection of labels is an important step for online learning when using less powerful computers. Although random selection may miss incorrectly-classified labels at a higher rate, it was experimentally verified that correct models can be obtained eventually. Therefore, we also adopted random selection.

1. Randomly select one relevant label r_t from Y_t.
2. Define irrelevant label s_t with random selection from Y_t and compute the hinge-loss $1 - (\boldsymbol{\mu}_t^{r_t} \cdot \boldsymbol{x}_t - \boldsymbol{\mu}_t^{s_t} \cdot \boldsymbol{x}_t)$. Continue selecting s_t until the loss becomes positive.
3. If the hinge-loss becomes positive, update models for r_t and s_t according to PA; otherwise move on to next training sample.

4 Results

In this section, we describe the details of methods we use for Flickr Photo annotion task, Web Photo subtask1, and Web Photo subtask2, respectively.

4.1 Photo Flickr

In our experiment, we extracted 5 kinds of visual descriptors from each image, SIFT and LBP in five patch sizes, and color-SIFTs (C-SIFT, RGB-SIFT, OpponentSIFT) in three patch sizes. As pre-processing, the images were resized into at most 300×300 pixels, of which aspect ratio were maintained. To calculate SIFT and LBP, the images were rendered in gray scale, in contrast to color-SIFTs which utilized color information. Then, each descriptor was densesampled on regular grids (every six pixels). The dimensionalities of SIFT, LBP, and color-SIFTs were 128, 1024, and 384 respectively. All of these descriptors were reduced to 64 dimensions with PCA, and then coded into two state-of-the-art global feature representations. ($5 \times 2 = 10$ visual features in total) One is FV, explained in the previous section. At first, we trained the mixture model of 256 Gaussians using standard EM-algorithm. To embed spatial information, FVs were calculated respectively over 1×1, 2×2, and 3×1 cells. In this way, we obtained FVs whose dimensionality was $64 \times 256 \times 8 \times 2 = 262,144$. The other is Locality-constrained Linear Coding (LLC) [11], which describes each local descriptor by a linear weighted sum of a few nearest codewords. In our experiment, 4,096 codewords were generated with k-means algorithm, and then each local descriptor was approximated using 3-NN of the descriptor. The images were divided into 1×1, 2×2 and 3×3 spatial grids differently from FV, so the dimensionality was $4096 \times 14 = 57,344$.

As text features, BoW vectors were extracted from Flickr-tags, and then we also prepared the one whose dimensions were removed if the corresponding words appeared 24 times or less. Furthermore, all combinations of 2 kinds of processing (TF-IDF, L_2-normalization) were done, so $2 \times 2^2 = 8$ text features were generated in total. Finally, the classifiers of 10 visual features and 8 text features were trained separately using PAAPL. The trade-off parameter was set to $C = 10^5$. All the experiments in the following sub-section were implemented using a workstation with CPUs of 12-core and 96GB RAM.

The validation set consisted of one-third of the training images, and validation was done only two times for the lack of time.

The size of the visual feature such as FV or LLC tends to be large. It is known to be effective to quantize the vector with Product Quantization (PQ) as described in [7, 4]. At first, we investigated the performance effect of PQ using FV-SIFT. The parameter of PQ was decided empirically, $b = 1$, $G = 8$. We iterated PAAPL learning 15 times. As a result, FV-SIFT achieved F1-measure (F1) 0.5604 with PQ while it achieved 0.5632 without PQ. Because the drop of performance was actually not significant, we quantized visual features of training samples for saving the RAM.

Since the number of runs which could be submitted was limited, we examined which combinations of visual features were effective. Then we investigated which text feature should be added to achieve the best performance through the next experiment.

The Table 1 shows the top six F1-measures of the combinations of $2^{10} = 1,024$ visual features. These features were all quantized with PQ. Note that all LLCs are shown to be inefficient here. In this way, we chosen to extract FVs from SIFT, from C-SIFT, and from LBP, which achieved the best performance.

FV-SIFT	✓	✓	✓	✓	✓	✓
FV-LBP	✓	✓	✓	✓	✓	✓
FV-OpponentSIFT	-	-	-	✓	-	✓
FV-cSIFT	✓	✓	-	-	✓	✓
FV-rgbSIFT	-	✓	✓	-	✓	-
LLC-SIFT	-	-	-	-	-	-
LLC-LBP	-	-	-	-	-	-
LLC-OpponentSIFT	-	-	-	-	-	-
LLC-cSIFT	-	-	-	-	-	-
LLC-rgbSIFT	-	-	-	-	✓	-
F1-measure	0.5715	0.5707	0.5703	0.5693	0.5693	0.5688

Table 1. Top combinations of visual features.

The Table 2 shows the F1-measures of eight text features. As a result, thresholding w.r.t. the number of corresponding images is not effective for the performance. Therefore, we chosen only four text features which were not thresholded.

threshold	histgram		TF-IDF	
		L_2		L_2
-	0.5157	0.5156	0.5135	0.5121
✓	0.5114	0.5109	0.5020	0.4990

Table 2. Top combinations of visual features.

Finally, we present the top six F1-measures (F1) of the combinations of three visual features and for text features in the Table 3. [2] Following these results, we submitted our runs described in Table 4 and got the scores also shown in Table 4. Note that all of these visual features from test images were not quantized.

[2] FV from SIFT is not quantized here.

Visual (FV)				Textual (BoW)			**F1**
FV-SIFT	FV-LBP	FV-cSIFT	histgram	TF-IDF	histgram (L_2)	TF-IDF (L_2)	
✓	✓	✓	✓	-	-	-	0.5798
✓	✓	-	✓	-	-	-	0.5792
✓	✓	-	-	-	✓	-	0.5770
✓	✓	✓	-	-	✓	-	0.5763
✓	✓	-	-	✓	-	-	0.5760
✓	✓	✓	✓	-	✓	-	0.5759

Table 3. Top combinations of visual and textual features. L_2 means the vector is normalized according to its L_2 norm.

method	**MiAP**	**GMiAP**	**F1**
FV-SIFT + FV-LBP	0.3243	0.2590	0.5451
FV-SIFT + FV-LBP + text	0.4046	0.3436	0.5559
FV-SIFT + FV-LBP + text(TF-IDF)	0.4029	0.3462	0.5597
FV-SIFT + FV-LBP + FV-C-SIFT + text	0.4136	0.3540	0.5574
FV-SIFT + FV-LBP + FV-C-SIFT + text(TF-IDF)	0.4131	0.3580	0.5583

Table 4. All five submissions and their scores on the Flickr Photo task. MiAP and GMiAP stand for (Geometric) Mean interpolated Average Precision

4.2 Photo Web

For Photo Web tasks, we could not extract FVs because there were no images in the provided dataset. Hence we use the provide BoVWs from some kinds of SIFTs.

Moreover, our PAAPL requires labels for each training sample. We investigated a simple way to define the labels for each training sample. In particular, we extracted words whichever are concept words from the surrounding texts for each image. Images around which any concepts do not exist are just discarded.

Subtask 1: Improving performance in Flickr concept annotation task For Flickr and Web image representation, we made use of four provided BoVWs, which were computed respectively from SIFT, C-SIFT, OpponentSIFT and RGB-SIFT. Unlike Flickr data, Web data was not annotated. Therefore we needed to estimate labels of web data for using it as training data in supervised learning. In order to do this, we sought for concepts in surrounding texts. If any concept labels exist in a textual feature, then we consider that the corresponding image has that label.

To annotate images, we use PAAPL with regularization parameter from 10^4, 10^5 and 10^6. As a result, $C = 10^6$ achieves the best performance in almost all cases where a classifier is trained on different type descriptors. The number of training iteration is 25, same as the Flickr Photo task.

We have two ideas on how to utilize web data for improvement of annotation performance. One idea is that we use Web data and Flickr data independently. At first, we trained eight classifiers using four types BoVWs from either Web data or Flickr data. Then, we summed the scores form the eight classifiers when we annotated the test images. To find best combinations of descriptors, we used 10000 Flickr images or 10000 Web images for training, and 5000 Flickr data for validation. We computed F1-measure as a measure of effectiveness of a combination. Results are shown in Table 5. The best F1-measure is worse than that

		Flickr			
		SIFT	C-SIFT	O-SIFT	RGB-SIFT
Web	SIFT	0.2086	0.2195	0.2162	0.2119
	C-SIFT	0.2158	0.2207	**0.2220**	0.2195
	O-SIFT	0.2190	**0.2276**	**0.2252**	**0.2227**
	RGB-SIFT	0.2009	0.2112	0.2062	0.2031

Table 5. Results of descriptor combinations. O-SIFT means OpponentSIFT.

obtained by the other idea, so we do not adopt this idea.

The other idea is that we merge 15000 Flickr data and 250000 web data , and unified 265000 data is used for training. Classifiers are trained respectively for each type of BoVW computed from different descriptor, so we have four classifiers. Scores for each label are computed by summing scores from each different classifiers. A number of ways of combining classifiers is $\sum_{i=1}^{4} {}_4\mathrm{C}_i = 15$. To find best combinations of four different classifiers, we use 10000 Flickr images and 10000 Web images for training, and 5000 Flickr data for validation. Then we computed F1-measure shown in Table 6.

Therefore we submitted the following combinations of BoVWs.

1. SIFT + C-SIFT
2. SIFT + C-SIFT + Opponent SIFT
3. SIFT + C-SIFT + RGB-SIFT
4. SIFT + C-SIFT + Opponent SIFT + RGB-SIFT

In combination of Web and Flickr photos, we obtained MiAP 0.264, GMiAP 0.217 and F1 0.182. However, we obtained MiAP 0.719, GMiAP 0.689 and F1 0.553 when we use only Flickr photos. This means that improving performance in Flickr concept annotation task using Web photos is not successful. Although improving performance in Web concept annotation task using Flickr photos seems to be meaningful, Web images are too noisy with such a simple way to combine.

Subtask 2: Scalable concept image annotation Before we achieved the results, we took three steps as follows.

SIFT	C-SIFT	O-SIFT	RGB-SIFT	**F1**
✓	-	-	-	0.4682
-	✓	-	-	0.4792
-	-	✓	-	0.4666
-	-	-	✓	0.4665
✓	✓	-	-	**0.4824**
✓	-	✓	-	0.4747
✓	-	-	✓	0.4738
-	✓	✓	-	0.4807
-	✓	-	✓	0.4817
-	-	✓	✓	0.4748
✓	✓	✓	-	**0.4833**
✓	✓	-	✓	**0.4821**
✓	-	✓	✓	0.4771
-	✓	✓	✓	0.4821
✓	✓	✓	✓	**0.4822**

Table 6. Results of combination experiments. O-SIFT means OpponentSIFT.

First, in order to assign concepts to each image, we compared the concepts to the raw text extracted near each image. If the raw text contains a concept, the concept is assigned as one of the concepts of the image.

Second, using randomly sampled 10000 training data from web and test data from all development set, we made grid search as for the two parameters in PA, that is $C = \{10^4, 10^5, 10^6\}$ and the iteration number $N = \{10, 15, 20, 25\}$. These candidates are empirically selected. As a result, shown in Fig.2, we chose $\{C_{C-SIFT} = 10^4, C_{OpponentSIFT} = 10^4, C_{RGB-SIFT} = 10^4, C_{SIFT} = 10^6\}$ and the iteration number $N = 25$.

Finally, utilizing the parameters stated above, we trained the weight vectors $\boldsymbol{\mu}$ corresponding to each feature from all 250k web data. After training, we examined all combinations ($2^4 = 16$) calculated by summing up the candidates from the four dot products of weight vectors and BoVW feature vectors. We assigned three concepts which had highest combined scores to each image from development set. We submitted five runs from 16 combinations as the results of the development set by calculating mean F1-measure (shown in Table 7) and used the same five combinations in order to achieve the results of the test set.

Therefore we submitted the following combinations of BoVWs.

1. C-SIFT + O-SIFT
2. C-SIFT + RGB-SIFT
3. O-SIFT + RGB-SIFT
4. C-SIFT + O-SIFT + RGB-SIFT
5. SIFT + C-SIFT + Opponent SIFT + RGB-SIFT

As a result, we obtained MiAP 0.332, GMiAP 0.227, and F1 0.254. These scores are higher than those we obtained with Web images in Subtask 1.

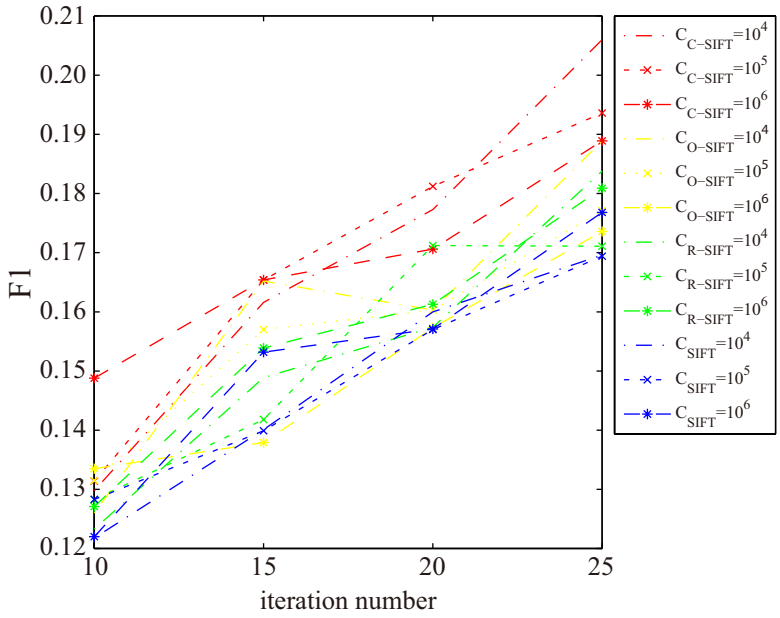

Fig. 2. Results of grid search on randomly selected 10k train data from web set and test data from development set as a function of iteration number and mean F1-measure (F1) for the four BoVW features.

5 Conclusions

In this working note, we describe our method to annotate images in ImageCLEF 2012 Photo Annotation and Retrieval tasks. We pay our attention to make our method scalable for a large amount of images. Consequently, we use FVs and BoWs in Photo Flickr task, and use the provided BoVW in Photo Web tasks. Annotation itself is achieved using a novel online learning PAAPL, which has been already proposed in [9] for multilabel problem.

For Photo Flickr task, we have achieved the top scores among all teams although our system is scalable and simple. Not only FVs from SIFTs but also FV from LBP is shown to be useful for annotation. Moreover, simple tag information with BoW improves the performance. For Photo Web tasks, there are few teams that have submitted at least one run. In Subtask 1, there are no teams that have improved the performance with Web data. The result of Subtask 2 also indicates that the Web Photoes are difficult to be extracted their proper concepts from their Web pages.

SIFT	C-SIFT	O-SIFT	RGB-SIFT	**F1**
✓	-	-	-	0.237
-	✓	-	-	0.258
-	-	✓	-	0.251
-	-	-	✓	0.244
✓	✓	-	-	0.257
✓	-	✓	-	0.256
✓	-	-	✓	0.247
-	✓	✓	-	**0.267**
-	✓	-	✓	**0.260**
-	-	✓	✓	**0.262**
✓	✓	✓	-	0.256
✓	✓	-	✓	0.258
✓	-	✓	✓	0.257
-	✓	✓	✓	**0.266**
✓	✓	✓	✓	**0.264**

Table 7. Results of summing up the combinations of the scores from each BoVW feature on train data from all web set and test data from all development set.

References

1. Bottou, L.: Large-Scale Machine Learning with Stochastic Gradient Descent. In: COMPSTAT (2010)
2. Crammer, K., Dekel, O., Keshet, J., Shalev-Shwartz, S., Singer, Y.: Online Passive-Aggressive Algorithms. JMLR 7, 551–585 (2006)
3. Grangier, D., Monay, F., Bengio, S.: A Discriminative Approach for the Retrieval of Images from Text Queries. In: ECML (2006)
4. Lin, Y., Lv, F., Zhu, S., Yang, M., Cour, T., Yu, K., Cao, L., Huang, T.: Large-scale Image Classification: Fast Feature Extraction and SVM Training. In: CVPR (2011)
5. Mensink, T., Csurka, G., Perronnin, F., Sanchez, J., Verbeek, J.: Lear and xrce's participation to visual concept detection task. In: CLEF 2010 working notes (2010)
6. Perronnin, F., Sánchez, J., Mensink, T.: Improving the Fisher Kernel for Large-Scale Image Classification. In: ECCV (2010)
7. Sánchez, J., Perronnin, F.: High-Dimensional Signature Compression for Large-Scale Image Classification. In: CVPR (2011)
8. Thomee, B., Popescu, A.: Overview of the imageclef 2012 flickr photo annotation and retrieval task. In: CLEF 2012 working notes (2012)
9. Ushiku, Y., Harada, T., Kuniyoshi, Y.: Efficient Image Annotation for Automatic Sentence Generation. In: ACM MM (2012, accepted)
10. Villegas, M., Paredes, R.: Overview of the imageclef 2012 scalable web image annotation task. In: CLEF 2012 working notes (2012)
11. Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., Gong, Y.: Locality-constrained Linear Coding for Image Classification. In: CVPR (2010)
12. Weston, J., Bengio, S., Usunier, N.: WSABIE: Scaling Up To Large Vocabulary Image Annotation. In: IJCAI (2011)