Visual Concept Detection and Annotation via Multiple Kernel Learning of Multiple Models

Yu Zhang[1], Stephane Bres[2], and Liming Chen[1]

[1] Universite de Lyon, CNRS, Ecole Centrale de Lyon,
LIRIS, UMR5205, F-69134, France
[2] LIRIS-INSA de Lyon, 20 avenue Albert Einstein, 69621 Villeurbanne Cedex, France
{Yu.Zhang,liming.chen}@ec-lyon.fr, stephane.bres@insa-lyon.fr

Abstract. This paper presents a multi-model framework for Visual Concept Detection and Annotation(VCDA) task based on Multiple Kernel Learning(MKL), To extract discriminative visual features and build visual kernels. Meanwhile the tags associated with images are used to build the textual kernels. Finally, in order to benefit from both visual models and textual models, fusion is carried out by MKL efficiently embed. Traditionally the term frequencies model is used to capture this useful textual information. However, the shortcoming in the term frequencies model lies in the fact that the performance seriously depends on the dictionary construction and in the fact that the valuable semantic information can not be captured. To solve this problem, we propose one textual feature construction approach based on $WordNet$ distance. The advantages of this approach are three-fold: (1) It is robust, because our feature construction approach does not depend on dictionary construction. (2) It can capture tags semantic information which is hardly described by the term frequencies model. (3) It efficiently fuses visual models and textual models. The experimental results on the ImageCLEF 2011 show that our approach effectively improves the recognition accuracy.

1 Introduction

The Visual Concept Detection and Annotation(VCDA) task is a multi-label classification challenge. The goal of this task is to decide whether a large number of images, which come from consumers, belongs to a certain concepts[5]. However the images coming from consumer include sense, events, or even sentiments. Due to large intra-class variations and inter-class similarities, clutter, occlusion and pose changes, this work is proved to be extremely challenging in computer vision domain.

State-of-the-art methods on VCDA mostly have focused on appropriate visual content descriptors and are still less capable of textual descriptor. Although tags associated with images from host or guest tend to be noisy in the sense that not directly relate to the image content, there is still much information in tags. This information is hard to describe by visual descriptor. Usually the term frequencies model is used to solve this problem. The tags are often represented

A. Petrosino (Ed.): ICIAP 2013, Part II, LNCS 8157, pp. 581–590, 2013.

as bag-of-words(BoW) model, each component of the vector is a kind of word count or term frequencies. The BoW approach achieves good performance on the VCDA task. This model has undergone several extensions, including latent semantic analysis(LSA), probabilistic latent semantic analysis(pLSA) and Latent Dirichlet allocation(LDA). However in this approach there are two drawbacks: (1) The BoW only considers the word frequency information, thus disregards tags semantic information. (2) The BoW is sensitive to the changes in dictionary that occur when training data can not be reasonably expected to be representative of all the potential testing data.

Recently in order to solve that the BoW only considers the word frequency information, disregards tags semantic information, Ningning Liu et al[7] propose that building textual feature based on $WordNet$ distance for VCDA task and demonstrate that it especially improves performance of VCDA task. However it is still seriously sensitive to the changes in dictionary.

In other hand, for visual information the VCDA task typically presents images with histograms or distribution of features from channels such as texture, color and local gradients[13]. This means that using only a single unified feature may not satisfactory solve the problem. In order to benefit from both visual models and textual models, the multiple kernel learning (MKL) approach carrys out the VCDA task with mix of ensemble the visual kernels and the textual kernels machines[6], as show in figure 1.

The main contributions of this work are summarized as follows:

- Building semantic textual feature. This approach can capture tags semantic information which is hardly described by the term frequencies models.
- Using $WordNet$-based semantic distance for feature construction. This approach is robust, because this method does not depend on dictionary construction.

In the next section we introduce our approach. Section 3 presents the proposed approach for VCDA, the experiment results are shown in section 4. Finally some conclusions are made in section 5.

2 Our Approach

2.1 Textual Models

Semantic Distance. We relay on the $WordNet$ to measure the distance between two words. $WordNet$ structure[4] can be seen as a semantic network where each node represents a concept of the real world. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. These synsets are connected by arcs that describe relations between concepts. The semantic distance between w_1 and w_2 is defined by:

$$distance(w_1, w_2) = \begin{cases} sim(s_1, s_2) & \text{if} \quad s_1, s_2 \ \exists \ CS \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

$$sim(s_1, s_2) = \frac{\min\{lcs(s_1), lcs(s_2)\}}{depth(CS) + \min\{lcs(s_1), lcs(s_2)\}} \quad (2)$$

Where s_i is a synset and $w_i \in s_i$. $lcs(s)$ denotes the distance form s to the common subsume (CS) (most specific ancestor node) of the two synsets s_1 and s_2 in a $WordNet$ taxonomy. $depth(CS)$ is the length of the path from CS to the taxonomy $Root$.

Semantic Textual Feature. Recently, Wang gang et al. [14] build textual feature for image object classification and demonstrate that it improves performance of visual object classification especially when the training dataset is small. By contrast, we propose a novel textual feature building approach for image classification, which is expected to capture part of the semantic meanings from images and more directly reflects the semantics of the scene in images. Meanwhile by computing the distance between tags set associated with images, the distance between images can be directly measured. The procedure for our approach is shown as Table 1. With this approach, it avoids relying on the construction of an dictionary.

Table 1. The procedure of the semantic textual feature building algorithm

Semantic textual feature

Input: Training dataset $Tr = \{Tr_1, Tr_2, \ldots, Tr_n\}$ and Testing dataset $Te = \{Te_1, Te_2, \ldots, Te_m\}$.
Output: The n-length feature vector f.

- preprocess the tags by using a stop-words filter.
- Build tags representation of Tr and Te data
  - For each $Te_i \in Te$ or $Tr_i \in Tr$
  - if Te_i or Tr_i has no tags, return $f_{ij} = 0$.
  - else Te_i or Tr_i has tags.
    - For each tags set $Tr_j \in Tr$
      - For each words $w_x \in Te_i$ or $w_x \in Tr_i$
      - For each words $w_y \in Tr_j$
      - $f_{ij} = f_{ij} + distance(w_x, w_y)$

Frequency Textual Feature. We consider here that the tag importance increases proportionally to the number of times a tag appears in the tags set of an image but is offset by the frequency of the tag in the corpus. when the tag set of an image is just a list of words, each tag appears just one. But sometimes, the tag set can be a text associated to an image. So we can employ the tf/idf approach to build the textual feature. In the dictionary we calculate the weight

of every tag. Finally we build the frequency textual feature. We calculate the tag weighting.

$$tf/idf = \frac{n_{i,j}}{\sum_k n_{k,j}} log \frac{|D|}{|j : t_i \in d_j|} \tag{3}$$

We define all tag set to be our corpus, j and compute the tf/idf score where n_{ij} represents the frequency of term i in the tag set of image i. The inverse tag frequency is computed as the log of the number of images $|D|$ divided by the number of tag set containing the term i.

2.2 Visual Models

Visual Features. Commonly the visual content of an image is described by visual descriptors such as color, texture, shape, etc. within a global or a bag of local features. In this work, we make use of several popular local descriptors,including C-SIFT, Rgb-SIFT, Hsv-SIFT, Oppo-SIFT and DAISY, extracted from a dense grid. Meanwhile, in order to capture the global ambiance and layout of an image, we further compute a set of global features, including descriptions of color information, in terms of LBP, Color LBP [18].

Bag-of-Features Representation. After local feature extraction, each input image is represented by a set of local descriptors. Because of the large number of sampling points (normally more than thousands), it is unreasonable to feed them directly into the classifier. Meanwhile these descriptors can not directly bridge the gap between visual descriptors and the semantic content of image. Therefore, we employ the dominant Bag-of-Features (BoF) method[2] which views an image as an unordered distribution of local image features extracted from dense image points[10] and transform these high dimensional descriptors to more compact and informative representations. The main idea of the BoF is to represent an image as an unorderless collection of local descriptors. More precisely, a visual vocabulary is constructed at first by applying a clustering algorithm such as k-means on the training data, and each cluster center is considered as a visual word in the vocabulary. All feature descriptors extracted from an image are then quantized to their closest visual word in an appropriate metric space. Finally the images are represented as fix-length vectors.

2.3 Multiple Kernels Learning

Due to the possibly large intraclass feature variations, using only a single unified kernel-based classifier may not satisfactorily solve the problem. Instead of selecting a single kernel, MKL learns a convex kernel combination and the associated classifier simultaneously; the combination of multi-kernels is defined as follows:

$$K(x_i, x) = \sum_{m=1}^{M} d_m K_m(x_i, x) \tag{4}$$

with $\sum_{m=1}^{M} d_m = 1$ and $d \geq 0 \ \forall m$ where M is the total number of kernels, $K_m = \phi_m(x_i)\phi_m(x_j)$ is a positive definite kernel which represents the dot product in feature space ϕ, and $\{d_m\}_{m=1}^{M}$ are kernel weights which are optimized during training. Each K_m can employ different kernel functions and use different feature subsets or data representations.

For binary classification, given the learning set $\{x_i, y_i\}_{i=1}^{M}$, where x_i belongs to some input data and y_i is the label of x_i, the decision function of canonical MKL is given as follows:

$$f(x) = \sum_{i=1}^{N} \alpha_i^* y_i \sum_{m=1}^{M} d_m K_m(x_i, x) + b^* \tag{5}$$

Where $\{\alpha_i^*\}_{i=1}^{N}$ and b^* are the coefficients of the classifier, corresponding to the lagrange multipliers and the bias in the canonical SVM problem. To solve the MKL problem efficiently, the SMO-MKL algorithm is used to optimise the l_p MKL dual[15].

3 The Proposed Approach for VCDA

Our framework for VCDA is depicted in Fig 1.

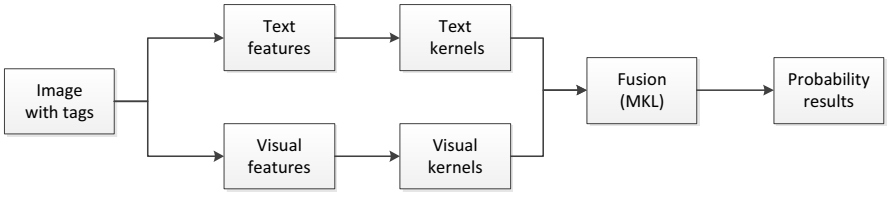

Fig. 1. The framework of our approach

3.1 Textual Model Construction

Semantic Textual Feature Construction. Our motivation of building textual feature based on $WordNet$ is to capture tags semantic information and eliminate the influence of the dictionary construction. In this work we restrict ourselves to the noun component of $WordNet$ and use only hyponymy and instance hyponymy relations for textual feature construction. After preprocessing and stemming, the process of suffix removal to generate word stems, Data set $D\{I_i, T_i\}$ consists of image I_i and tags set T_i. The weight between I_i and I_j are measured by T_i and T_j. We compute the distance between each word of tag set T_i and each word of tag set T_j and each word of tag set T_j according to Function 1. The overview of the experiment procedure is shown as table 1.

Frequency Textual Feature Construction. The tf/idf feature is employed to capture tag frequency information. The tags that appear at least 3 times (a minimum of 3 times in the training set) are used as the dictionary, resulting in a dictionary of 5154 words in the data set of ImageCLEF 2011, which is the one we use for our test here. Finally each image is represented by a BoW histogram of 5154 dimensions.

3.2 Visual Model Construction

Note that in our paper, features come from multiple sources. Visual features include color-SIFT, color-LBP and DAISY, which are used to capture image content from channels such texture, color and local gradients[9]. For global feature color LBP, multi-scale color LBP descriptors based on scale 8, 12 and 16 are employed. For local features color SIFT and DAISY, the sampling spacing is set to 6 pixels. A visual vocabulary with 4000 visual words is then constructed by applying k-means clustering algorithm to 800,000 randomly selected descriptors from the training set. Each image is finally transformed to fixed-length features.

3.3 Fusion and Classification

The chi-square kernel(χ^2 distance) is used to measure the similarity between two feature vectors F and F' (n is the size of the feature vector). Then, the kernel function based on this distance is used for MKL to train the classifier:

$$K_{\chi^2}(F, F') = e^{-\frac{1}{D}\sum_{i-1}^{n}\frac{(F_i - F_i')^2}{F_i + F_i'}} \tag{6}$$

Where D is the parameter for normalizing the distances. Here D is set to the average distance of all the training data. Once giving kernels, MKL seeks to the best combination-weights of these kernels.

4 Experimental Evaluation

In our experiment the ImageCLEF 2011 dataset with 99 concepts are employed. The training set consists of 8000 photos, and the testing set consists of 10000 photos. All photos are associated with EXIF data and Flickr user tags, For evaluation, we use mean average precision (mAP)[17].

4.1 Results: Visual Models

We apply different types of visual features to build the visual models and fuse same types of visual features with MKL respectively on ImageClEF 2011 dataset. The experimental results of each single visual feature and fusion approach are shown in Fig.2. For each single visual feature, we can see that the color SIFT based features outperform other descriptors. The performances of color SIFT

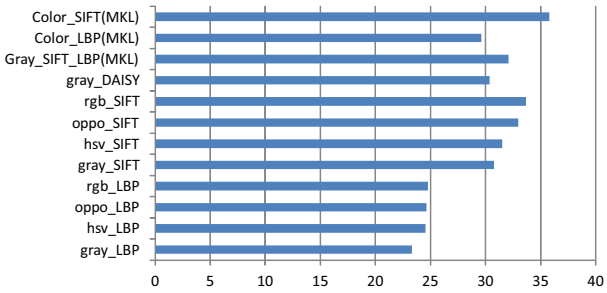

Fig. 2. The mAP performance of different visual models

features obtain about 30% $\sim$ 34% mAP value. Moreover compared with single visual feature, the performance of multi-visual model is better.

Table 2 shows the performance of different teams who participated the ImageCLEF 2011 challenge. TUBFI's, CAEN's, ISIS's and BPACAD's visual model ranked the $1^{st}, 2^{nd}, 3^{rd}$ and 4^{th}. Compared with their results, the performance of our visual model is comparable.

Table 2. Comparison of our visual model with other's on ImageCLEF 2011

Teams(Visual model)	mAP (%)
TUBFI[1]	38.8
CAEN[14]	38.2
ISIS[12]	37.5
BPACAD[3]	36.7
Color_LBP_SIFT(MKL)	37.4

4.2 Results: Textual Models

We compare Term Frequency, TF/IDF, LDA and HTC[8] approach with the proposed semantic textual feature. The mAP performances are shown in table 3. The results indicate that the performance of our textual model is not good. The main reason may be that our approach only considers the tags sematic relation, compared with other textual approaches. Moreover, in order to capture the frequency and semantic information, we employ MKL approach to fuse semantic textual model and frequency model. We can see our multi-textual model outperforms other methods almost 5% in mAP evaluation.

Finally, in order to evaluate our approach, we compare our approach with the textual configuration results which are obtained top 4 in ImageCLEF 2011 challenge. It can be seen that the semantic BoW model outperforms all team's results.

Table 3. Comparison of different textual models on ImageCLEF 2011

Textual model	dictionary size	mAP (%)
Term Frequency	5154	32.53
tf/idf	5154	32.41
LDA	2500	31.35
HTC	2000	32.12
semantic textual feature	-	27.15
tf/idf_semantic(MKL)	-	37.48

Table 4. Comparison of our textual model with other's on ImageCLEF 2011

Teams(Textual model)	mAP (%)
BPACAD	34.6
IDMT[11]	32.6
MLKD[16]	32.6
LIRIS[8]	32.1
tf/idf_semantic(MKL)	37.5

4.3 Results: Fusion of Visual Models and Textual Models

The MKL approach is employed to fuse the textual model and the visual model. The different types of visual models are fused with textual models. The experimental results are shown in table 5. The results notices that combining multiple feature channels can improve the performances. Meanwhile we investigated the results of TUBFI, Liris, BPACAD, ISIS and MLKD, whose multi-model approaches ranked in top 5 of the challenge 2011 on mAP evaluation, as shown in table 6. TUBFI applied non-sparse multiple kernel learning and multi-task learning to build classifiers. To build the textual features, they used BoW and Markov random walks based on the Flickr user tags. Compared with other team's results, our approach gets the best result of 45.73% mAP.

Table 5. The mAP performance of different multi-model approach on ImageCLEF 2011

Multi model(MKL)	mAP (%)
LBP_Text	42.26
SIFT_Text	44.24
LBP_SIFT_Text	45.73

Fig 3 shows a part of the Average Precision per concept in detail, and it can be noticed that our results significantly outperform the TUBFI's best run on the concepts of airplane and skateboard. Analysis shows that the number of training samples for these concepts are only 41 and 12, which makes it extremely difficult to classify those concepts. However, our textual features improve the performance of our visual classifiers regarding to these cases.

Table 6. Comparison of our multi-model with other's on ImageCLEF 2011

Teams(Multi model)	mAP (%)
TUBFI	44.3
LIRIS	43.7
BPACAD	43.6
ISIS	43.3
MLKD	40.2
LBP_SIFT_Text(MKL)	45.7

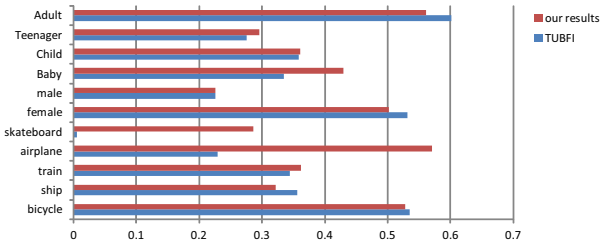

Fig. 3. A part of the Average Precision per concept of our best multi-model runs compared to TUBFI's

5 Conclusion

In this paper, we focused on the problem of how the tags associated with images can benefit for automatic visual concept detection and annotation. We proposed a novel method to build textual descriptor based on the semantic distance between the user tags. Meanwhile we introduced a novel multi-model approach with MKL for the VCDA task. The main contributions are that the semantic textual feature can easily capture semantic information contained in tags which is hardly described by the term frequencies model. Comprehensive experiments were conducted on the ImageCLEF 2011 dataset. Compared with the other approaches, our approach exhibits the best preferences. From the experiment results, we conclude the following: (1) Based on proposed the approach, it consistently improves the performance of visual classifiers, especially when the concept training set is small. (2) The multi-model approach is especially useful for the VCDA task with multi-label scenario.

References

1. Binder, A., Samek, W., Kloft, M., Müller, C., Müller, K.R., Kawanabe, M.: The joint submission of the tu berlin and fraunhofer first (tubfi) to the imageclef 2011 photo annotation task. In: Petras, V., Forner, P., Clough, P.D. (eds.) CLEF (2011)
2. Csurka, G., Dance, C.R., Fan, L., Willamowski, J., Bray, C., Bray, C.: Visual categorization with bags of keypoints. In: European Conference on Computer Vision (ECCV), pp. 1–22 (2004)

3. Daróczy, B., Pethes, R., Benczúr, A.A.: Sztaki @ imageclef 2011. In: Petras, V., Forner, P., Clough, P.D. (eds.) CLEF (Notebook Papers/Labs/Workshop) (2011)
4. Fellbaum, C. (ed.): WordNet: an electronic lexical database. MIT Press (1998)
5. Guillaumin, M., Verbeek, J., Schmid, C.: Multimodal semi-supervised learning for image classification. In: CVPR (June 2010)
6. Lin, Y.Y., Liu, T.L., Fuh, C.S.: Local ensemble kernel learning for object category recognition. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2007)
7. Liu, N., Dellandréa, E., Tellez, B., Chen, L.: Associating textual features with visual ones to improve affective image classification. In: D'Mello, S., Graesser, A., Schuller, B., Martin, J.-C. (eds.) ACII 2011, Part I. LNCS, vol. 6974, pp. 195–204. Springer, Heidelberg (2011)
8. Liu, N., Zhang, Y., Dellandréa, E., Bres, S., Chen, L.: Liris-imagine at imageclef 2011 photo annotation task. In: Petras, V., Forner, P., Clough, P.D. (eds.) CLEF (Notebook Papers/Labs/Workshop) (2011)
9. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vision 60(2), 91–110 (2004), http://dx.doi.org/10.1023/B:VISI.0000029664.99615.94
10. Mikolajczyk, K., Schmid, C.: Indexing based on scale invariant interest points (2001), http://perception.inrialpes.fr/Publications/2001/MS01a
11. Nagel, K., Nowak, S., Kühhirt, U., Wolter, K.: The fraunhofer idmt at imageclef 2011 photo annotation task. In: Petras, V., Forner, P., Clough, P.D. (eds.) CLEF (Notebook Papers/Labs/Workshop) (2011)
12. van de Sande, K.E.A., Snoek, C.G.M.: The university of amsterdam's concept detection system at imageclef 2011. In: Petras, V., Forner, P., Clough, P.D. (eds.) CLEF (Notebook Papers/Labs/Workshop) (2011)
13. Siddiquie, B., Vitaladevuni, S.N.P., Davis, L.S.: Combining multiple kernels for efficient image classification. In: WACV, pp. 1–8. IEEE Computer Society (2009)
14. Su, Y., Jurie, F.: Semantic contexts and fisher vectors for the imageclef 2011 photo annotation task. In: Petras, V., Forner, P., Clough, P.D. (eds.) CLEF (Notebook Papers/Labs/Workshop) (2011)
15. Vishwanathan, S.V.N., Sun, Z., Theera-Ampornpunt, N., Varma, M.: Multiple kernel learning and the SMO algorithm. In: Advances in Neural Information Processing Systems (December 2010)
16. Xioufis, E.S., Sechidis, K., Tsoumakas, G., Vlahavas, I.P.: Mlkd's participation at the clef 2011 photo annotation and concept-based retrieval tasks. In: Petras, V., Forner, P., Clough, P.D. (eds.) CLEF (Notebook Papers/Labs/Workshop) (2011)
17. Yue, Y., Finley, T., Radlinski, F., Joachims, T.: A support vector method for optimizing average precision. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2007, pp. 271–278. ACM, New York (2007)
18. Zhu, C., Bichot, C.E., Chen, L.: Multi-scale Color Local Binary Patterns for Visual Object Classes Recognition. In: International Conference on Pattern Recognition (ICPR), pp. 3065–3068. IEEE (August 2010)