LIRIS-Imagine at ImageCLEF 2012 Photo Annotation task

Ningning Liu, Emmanuel Dellandréa, Liming Chen, Aliaksandr Trus, Chao Zhu, Yu Zhang, Charles-Edmond Bichot, Stéphane Bres, and Bruno Tellez

Université de Lyon, CNRS, Ecole Centrale de Lyon,
LIRIS, UMR5205, F-69134, France
{ningning.liu,emmanuel.dellandrea,liming.chen,aliaksandr.trus,chao.zhu,
yu.zhang,charles-edmond.bichot}@ec-lyon.fr
stephane.bres@insa-lyon.fr
bruno.tellez@univ-lyon1.fr
http://liris.cnrs.fr/

Abstract. In this paper, we present the methods we have proposed and evaluated through the ImageCLEF 2012 Photo Annotation task. More precisely, we have proposed the Histogram of Textual Concepts (HTC) textual feature to capture the relatedness of semantic concepts. In contrast to term frequency-based text representations mostly used for visual concept detection and annotation, HTC relies on the semantic similarity between the user tags and a concept dictionary. Moreover, a Selective Weighted Late Fusion (SWLF) is introduced to combine multiple sources of information which by iteratively selecting and weighting the best features for each concept at hand to be classified. The results have shown that the combination of our HTC feature with visual features through SWLF can improve the performance significantly. Our best model, which is a late fusion of textual and visual features, achieved a MiAP (Mean interpolated Average Precision) of 43.67% and ranked first out of the 80 submitted runs.

Keywords: textual features, visual feature, feature fusion, concept detection, photo annotation, multimodality, ImageCLEF

1 Introduction

Machine-based recognition of visual concepts aims at recognizing automatically from images high-level semantic concepts (HLSC), including scenes (indoor, outdoor, landscape, *etc.*), objects (car, animal, person, *etc.*), events (travel, work, *etc.*), or even emotions (melancholic, happy, *etc.*). It proves to be extremely challenging because of large intra-class variations (clutter, occlusion, pose changes, *etc.*) and inter-class similarities [1–4]. The past decade has witnessed tremendous efforts from the research communities as testified the multiple challenges in the field, *e.g.*, ImageCLEF [5–8], TRECVID [9] and Pascal VOC [10]. Increasing works in the literature have discovered the wealth of semantic meanings conveyed by the abundant textual captions associated with images [11–13]. As a

result, multimodal approaches have been increasingly proposed visual concept detection and annotation task (VCDT) by making joint use of user textual tags and visual descriptions to bridge the gap between low-level visual features and HLSC [7].

The VCDT is a multi-label classification challenge. It aims at the automatic annotation of a large number of consumer photos with multiple annotations. There were remarkable works have been proposed for ImageCLEF photo annotation tasks. The LEAR and XRCE group [14] in ImageCLEF 2010 employed the Fisher vector image representation with the TagProp method for image auto-annotation. The TUBFI group [15] in ImageCLEF 2011 built textual features using a soft mapping of textual Bag-of-Words (BoW) and Markov random walks based on frequent Flickr user tags. Our group in ImageCLEF 2011 [16] firstly proposed a novel textual representation, named Histogram of Textual Concept (HTC), which captures the relatedness of semantic concepts. Meanwhile we also proposed a novel selective weighted late fusion (SWLF) method, which automatically selects and weights the best discriminative features for each visual concept to be predicted in optimizing the overall mean average precision. This year, we have improved our approaches in the following aspects:

- We evaluated different textual preprocessing methods, and proposed enhanced HTC features using term frequency information. Meanwhile, we implemented two types of distributional term representations: documents occurrence representation (DOR) and DOR_TFIDF [17].
- We investigated a set of mid-level features, which are related to harmony, dynamism, aesthetic quality, emotional color representation, *etc.*. Meanwhile, we improved the harmony and dynamism features by adding a local information.

The rest of this paper is organized as follows. The features are introduced in Section 2, including textual and visual features as well as the fusion scheme proposed to combine them. The results are analysed in Section 3. Finally, Section 4 draws the conclusion and gives some hints for future work.

2 Features for semantic concepts recognition

In this section, we firstly present the textual features including HTC and enhanced HTC in Section 2.1, following with (Section 2.2) description of visual features which can be categorized into four groups: color, texture, shape and mid-level. The feature fusion scheme, SWLF, is presented in Section 2.3.

2.1 Textual features

The Histogram of Textual Concepts, HTC, of a text document is defined as a histogram based on a vocabulary or dictionary where each bin of this histogram represents a concept of the dictionary, whereas its value is the accumulation of

the contribution of each word within the text document toward the underlying concept according to a predefined semantic similarity measure.

The advantages of HTC are multiple. First, for a sparse text document as image tags, HTC offers a smooth description of the semantic relatedness of user tags over a set of textual concepts defined within the dictionary. More importantly, in the case of polysemy, HTC helps disambiguate textual concepts according to the context. For instance, the concept of "bank" can refer to a financial intermediary but also to the shoreline of a river. However, when a tag "bank" comes with a photo showing a financial institution, correlated tags such as "finance", "building", "money", *etc.*, are very likely to be used, thereby clearly distinguishing the concept "bank" in finance from that of a river where correlated tags can be "water", "boat", "river", *etc.* Similarly, in the case of synonyms, the HTC will reinforce the concept related to the synonym as far as the semantic similarity measurement takes into account the phenomenon of synonyms. The algorithm for the extraction of a HTC feature is detailed in the following algorithm:

The Histogram of Textual Concepts (HTC) Algorithm:

Input: Tag data $W = \{w_t\}$ with $t \in [1, \mathtt{T}]$, dictionary $D = \{d_i\}$ with $i \in [1, \mathtt{d}]$.
Output: Histogram f composed of values f_i with $0 \leq f_i \leq 1$, $i \in [1, \mathtt{d}]$.

- Preprocess the tags by using a stop-words filter.
- If the input image has no tags ($W = \emptyset$), return f with $\forall i \ f_i = 0.5$.[1]
- Do for each word $w_t \in W$:
  1. Calculate $dist(w_t, d_i)$, where $dist$ is a semantic similarity distance between w_t and d_i.
  2. Obtain the semantic matrix S as: $S(t, i) = dist(w_t, d_i)$.
- Calculate the feature f as: $f_i = \sum_{t=1}^{\mathtt{T}} S(t, i)$, and normalize it to [0 1] as: $f_i = f_i / \sum_{j=1}^{\mathtt{d}} f_j$.

---

[1] When an input image has no tag at all, in this work we simply assume that every bin value is 0.5, therefore at halfway between a semantic similarity measurement 0 (no relationship at all with the corresponding concept in the dictionary) and 1 (full similarity with the corresponding concept in the dictionary). Alternatively, we can also set these values to the mean of HTCs over the captioned images of a training set.

The computation of HTC requires the definition of a dictionary and a proper semantic relatedness measurement over textual concepts. For the ImageCLEF 2012 photo annotation task, we used two types of dictionaries. The first one is dictionary based on the term frequency on the training set, e.g. dictionary *TF_10T* consists of top 10 thousand words sorted by their frequencies in the training set. While the second one, *D_Anew*, is the set of 1034 English words used in the ANEW study [18]. The interest of the ANEW dictionary lies in the fact that each of its word is rated on a scale from 1 to 9 using affective norms

Table 1. The summary of textual features.

Short name	Description	
1	txtFtr_DOR	implement features of documents occurrence representation [17].
2	txtFtr_DOR_TFIDF	
3	txtFtr_HTC_Danew	obtained by using WordNet path distance on ANEW dictionary.
4	txtFtr_TFIDF_Danew	obtained on ANEW dictionary.
5	txtFtr_eHTC_Danew	obtained by adding each bins of txtFtr_4 and txtFtr_5.
6	txtFtr_TFIDF_TF_10T	obtained on the dictionary TF_10T, which is the top 10 thousand words sorted by the term frequency.
7	txtFtr_HTC_VAD	obtained using Eq. 1, Eq. 2 and Eq. 3.
8	txtFtr_HTC_TF_10T	obtained by using WordNet path distance on TF_10T dictionary.
9	txtFtr_HTC_TF_20T	obtained by using WordNet path distance on TF_20T dictionary.
10	txtFtr_TFIDF_TF_20T	obtained on TF_20T dictionary.
11	txtFtr_eHTC_TF_20T	obtained by adding each bins of txtFtr_9 and txtFtr_10.

in terms of *valence* (affective dimension expressing positive versus negative), *arousal* (affective dimension expressing active versus inactive) and *dominance* (affective dimension expressing dominated versus in control). For instance, according to ANEW, the concept "beauty" has a mean valence of 7.82, a mean arousal of 4.95 and a mean dominance of 5.23 while the concept "bird" would have a mean valence of 7.27, a mean arousal of 3.17 and a mean dominance of 4.42. Using the affective ratings of the ANEW concepts and the HTCs computed over image tags, one can further define the coordinates of an image caption in the three dimensional affective space [19], in terms of valence, arousal and dominance by taking a linear combination of the ANEW concepts weighted by the corresponding HTC values. More precisely, given a HTC descriptor f extracted from a text document, the valence, arousal and dominance coordinates of the text document can be computed as follows:

$$f_{valence} = (1/\mathtt{d}) \sum_i (f_i * V_i) \tag{1}$$

$$f_{arousal} = (1/\mathtt{d}) \sum_i (f_i * A_i) \tag{2}$$

$$f_{dominance} = (1/\mathtt{d}) \sum_i (f_i * D_i) \tag{3}$$

where V_i, A_i and D_i are respectively the valence, the arousal and the dominance of the i^{th} word w_i in the *D_Anew* dictionary, and $\mathtt{d}$ is the size of *D_Anew*.

The HTC features fail to calculate the semantic distance of two terms when the semantic relatedness measurement are not defined between these two terms. In order to cope with this problem, we enhanced the HTC features by combining it with TF/IDF features in a simple way: sum the value on each bin, and then normalize for the same dictionary. Meanwhile, we employed the distributional term representation DOR and DOR-TF/IDF [17]. A summary of textual features is given in Table 1.

2.2 Visual features

For ImageCLEF 2011 photo annotation task, we have introduced various visual features to describe interesting details and to catch the global image atmosphere. Thus, 5 groups of features have been considered: color, texture, shape, local descriptor and mid-level features [16]. This year, we have enriched this set of visual features by adding color SIFT features with 4000 codewords and soft assignment [20] and TOPSURF feature [21]. Moreover, we have enhanced the mid-level features harmony and dynamism by adding a local information through their computation using a pyramid grid.

2.3 Feature fusion through SWLF

In order to combined efficiently textual and visual features, we have proposed a Selective Weighted Late Fusion (SWLF) scheme which learns to automatically select and weight the best features for each visual concept to be recognized.

SWLF scheme has a learning phase which requires a training dataset for the selection of the best experts and their corresponding weights for each visual concept. Specifically, given a training dataset, we divide it into two disjoint parts composed of a training set and a validation set. For each visual concept, a binary classifier (concept versus no concept) is trained, which is also called expert in the subsequent, for each type of features using the data in the training set. Thus, for each concept, we generate as many experts as the number of different types of features. The quality of each expert can then be evaluated through a quality metric using the data in the validation set. In this work, the quality metric is chosen to be the interpolated Average Precision (iAP). The higher iAP is for a given expert, the more weight should be given to the score delivered by that expert for the late fusion. This fusion is performed as the sum of the weighted scores. More details on SWLF can be found in [22].

3 Experiments and Results

Our methods have been evaluated through the ImageCLEF 2012 photo annotation task, and particularly through the visual concept detection, annotation and retrieval subtask whose details are provided in [23]. There are 94 concepts

to automatically detect, that can be categorized into 5 groups: natural elements (day, night, sunrise, etc.), environment (desert, coast, landscape, etc.), people (baby, child, teenager, etc.), image elements (in focus, city life, active, etc.), human elements (rail vehicle, water vehicle, air vehicle, etc.).

In order to obtain a stable and better performance, we divided the training set into a training part (50%, 7501 images) and a validation part (50%, 7499 images) as required by SWLF presented in section 2.3.

3.1 The submitted runs

We submitted 5 runs to the ImageCLEF 2012 photo annotation challenge (2 textual model, 1 visual model and 2 multimodal models). All runs were based on the features described in the previous sections, including 11 textual ones and 32 visual ones. For the example evaluation, we propose two methods to chose the threshold. One is based on the distribution of training data. More specifically, we firstly calculate the distribution of concepts on the training set, then for each concept, we set the threshold as the boundary which makes the proportion of positive sample as same as it is in the training data. This idea is that we consider the training and test set share the same distribution for each concept. The other is to select a best threshold, which receives the best FMeasure value on the validation set. Based on the previous experiments and observations, we performed our runs based on the following configuration:

1. **textual_model_1**: the combination of the top 4 features among the 11 textual features for each concept based on the weighted score SWFL scheme.
2. **textual_model_2**: the combination of the top 6 features among the 11 textual features for each concept based on the weighted score SWFL scheme.
3. **visual_model_3**: the combination of the top 5 features among the 24 visual features for each concept based on the weighted score SWFL scheme.
4. **multimodal_model_4**: the combination of the top 22 features among the 43 visual and textual features for each concept based on the weighted score SWFL scheme.
5. **multimodal_model_5**: the combination of the top 26 features among the 43 visual and textual features for each concept based on the weighted score SWFL scheme.

3.2 Results

The results obtained by our 5 runs are given in Table 2. The best performance was provided by our multimodal models which outperformed the purely textual and purely visual ones. Moreover, our best model obtained the first rank based on the MiAP among the 80 runs submitted to the challenge.

Table 2. The results of our submitted runs.

Submitted runs	mAP(%)	GMiAP(%)	F-ex(%)
text_model_1	33.28	27.71	39.17
text_model_2	33.38	27.59	46.91
visual_model_3	34.81	28.58	54.37
multimodal_model_4	43.66	38.75	57.63
multimodal_model_5	**43.67**	38.77	57.66

3.3 Discussion

For the textual features, we proposed to apply two preprocessing methods. One is the removing of stopping words. The other one is stemming on 4 language (English, Germany, French, Italian). Based on the ImageCLEF 2012 photo annotation dataset, we find that after these two preprocessing, the MiAP performance of term frequency features e.g. TF/IDF, DOR improves about 1%. But the stemming is not proper for HTC features as it fails to calculate the semantic similarity measurement after stemming.

For the visual features, the harmony and dynamism features computed locally using a pyramid grid achieved 3% improvement on MiAP compared to the original ones.

For the HTC, we tested several semantic distances methods of WordNet including path, wup and lin. It is found that the path distance obtained the best performance.

4 Conclusion

We have presented in this paper the models that we have evaluated through the ImageCLEF 2012 photo annotation challenge. Our best multimodal prediction model which relies on the fusion through SWLF of our textual features (HTC) and visual features including low-level and mid-level information achieved a MiAP of 43.6% and ranked the best performance out of the 80 submitted runs. From the experimental results, we can conclude the following: (i) the proposed multimodal approach greatly improve the performance of purely textual and purely visual ones, with about 9% higher than the best visual-only model; (ii) the fused experts through weighted score-based SWLF, display a very good generalization skill on unseen test data and prove particularly useful for the image annotation task with multi-label scenarios in efficiently fusing visual and textual features.

In our future work, we envisage further investigation of the interplay between textual and visual content, in studying in particular the visual relatedness in regard to textual concepts. We also want to study some mid-level visual features or representations, for instance using an attentional model, which better account for affect related concepts.

Acknowledgement

This work was supported in part by the French research agency ANR through the VideoSense project under the grant 2009 CORD 026 02.

References

1. A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, R. Jain, Content-based image retrieval at the end of the early years, IEEE Trans. Pattern Anal. Mach. Intell. 22 (2000) 1349–1380.
2. A. Mojsilović, J. Gomes, B. Rogowitz, Semantic-friendly indexing and quering of images based on the extraction of the objective semantic cues, Int. J. Comput. Vision 56 (2004) 79–107.
3. J. Li, J. Z. Wang, Automatic linguistic indexing of pictures by a statistical modeling approach, IEEE Trans. Pattern Anal. Mach. Intell. (2003) 1075–1088.
4. M. S. Lew, N. Sebe, C. Djeraba, R. Jain, Content-based multimedia information retrieval: State of the art and challenges, TOMCCAP (2006) 1–19.
5. M. J. Huiskes, M. S. Lew, M. S. Lew, The mir flickr retrieval evaluation, in: Multimedia Information Retrieval, 2008, pp. 39–43.
6. M. J. Huiskes, B. Thomee, M. S. Lew, New trends and ideas in visual concept detection: The mir flickr retrieval evaluation initiative, in: MIR '10: Proceedings of the 2010 ACM International Conference on Multimedia Information Retrieval, 2010, pp. 527–536.
7. S. Nowak, K. Nagel, J. Liebetrau, The clef 2011 photo annotation and concept-based retrieval tasks, in: CLEF Workshop Notebook Paper, 2011.
8. S. Nowak, M. J. Huiskes, New strategies for image annotation: Overview of the photo annotation task at imageclef 2010, in: CLEF Workshop Notebook Paper, 2010.
9. A. F. Smeaton, P. Over, W. Kraaij, Evaluation campaigns and trecvid, in: MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval, 2006, pp. 321–330.
10. M. Everingham, L. J. V. Gool, C. K. I. Williams, J. M. Winn, A. Zisserman, The pascal visual object classes (voc) challenge, Int. J. Comput. Vision (2010) 303–338.
11. G. Wang, D. Hoiem, D. A. Forsyth, Building text features for object image classification., in: CVPR, 2009, pp. 1367–1374.
12. J. Sivic, A. Zisserman, Video google: A text retrieval approach to object matching in videos, in: ICCV, 2003, pp. 1470–1477.
13. M. Guillaumin, J. J. Verbeek, C. Schmid, Multimodal semi-supervised learning for image classification., in: CVPR, 2010, pp. 902–909.
14. T. Mensink, G. Csurka, F. Perronnin, J. Snchez, J. J. Verbeek, Lear and xrce's participation to visual concept detection task - imageclef 2010, in: CLEF Workshop Notebook Paper, 2010.
15. A. Binder, W. Samek, M. Kloft, C. Müller, K.-R. Müller, M. Kawanabe, The joint submission of the tu berlin and fraunhofer first (tubfi) to the imageclef2011 photo annotation task, in: CLEF Workshop Notebook Paper, 2011.
16. N. Liu, E. Dellandréa, L. Chen, Y. Zhang, C. Zhu, C.-E. Bichot, S. Bres, B. Tellez, LIRIS-Imagine at ImageCLEF 2011 Photo Annotation task, in: CLEF Workshop Notebook Paper, 2011.

17. H. J. Escalante, M. Montes, E. Sucar, Multimodal indexing based on semantic cohesion for image retrieval, Information Retrieval 15 (2011) 1–32.
18. M. M. Bradley, P. J. Lang, Affective norms for english words (ANEW): Stimuli, instruction manual, and affective ratings, Tech. rep., Center for Research in Psychophysiology, University of Florida, Gainesville, Florida (1999).
19. K. Scherer, Appraisal Processes in Emotion: Theory, Methods, Research (Series in Affective Science), Oxford University Press, USA, 2001.
20. J. C. van Gemert, C. J. Veenman, A. W. M. Smeulders, J. M. Geusebroek, Visual word ambiguity, IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (7) (2010) 1271–1283.
21. B. Thomee, E. M. Bakker, M. S. Lew, Top-surf: a visual words toolkit, in: Proceedings of the international conference on Multimedia, 2010, pp. 1473–1476.
22. N. Liu, E. Dellandrea, C. Zhu, C.-E. Bichot, L. Chen, A selective weighted late fusion for visual concept recognition, in: ECCV 2012 Workshop on Information fusion in Computer Vision for Concept Recognition, 2012.
23. B. Thomee, A. Popescu, Overview of the imageclef 2012 flickr photo annotation and retrieval task, in: CLEF 2012 working notes, Rome, Italy, 2012.