

LIRIS-Imagine at ImageCLEF 2011 Photo Annotation task

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

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

Abstract. In this paper, we focus on one of the ImageCLEF tasks that LIRIS-Imagine research group participated: visual concept detection and annotation. For this task, we firstly propose two kinds of textual features to extract semantic meanings from text associated to images: one is based on semantic distance matrix between the text and a semantic dictionary, and the other one carries the valence and arousal meanings by making use of the Affective Norms for English Words (ANEW) dataset. Meanwhile, we investigate efficiency of different visual features including color, texture, shape, high level features, and we test four fusion methods to combine various features to improve the performance including min, max, mean and score. The results have shown that combination of our textural features and visual features can improve the performance significantly.

Keywords: textual feature, visual feature, ImageCLEF photo annotation, multi-model, combination

1 Introduction

In our participation to the ImageCLEF visual concept detection and annotation task, we mainly focus on two aspects. Firstly, we investigated the effect of various descriptors representing semantic concepts in images including 10 novel textual features (Flickr-tags) [1, 2] and 24 visual features. Then, we studied different combination techniques to improve the performance of fusion of textual and visual models

The visual concept detection and annotation task is a multi-label classification challenge [6, 7]. It aims at the automatic annotation of a large number of consumer photos with multiple annotations. In 2011, the training set for annotation task consists of 8000 photos annotated with 99 visual concepts (93 concepts in 2010), and the testing set consists of 10000 photos with EXIF data and Flickr

user tags. These 99 concepts include the scene categories (indoor, outdoor, landscape, etc.), depicted objects (car, animal, person, etc.), the representation of image content (portrait, graffiti, art, etc.), events (travel, work, etc.) or quality issues (overexposed, underexposed, blurry, etc.). In this year, a special focus is laid to the detection of sentiment concepts by introducing Russell’s emotion model (8 concepts + ”funny” which is not included in this model). From 2010, the challenge has provided multi-model approaches that consider visual information and/or Flickr user tags and/or EXIF information. Thus, this task can be solved by following three different approaches [6, 7]:

- Automatic annotation with visual information only.
- Automatic annotation with Flickr user tags (tag enrichment).
- Multi-modal approaches that consider visual information and/or Flickr user tags and/or EXIF information.

We have investigated the remarkable works from ImageCLEF2010 photo annotation tasks. LEAR and XRCE group [12] employed the Fisher vector image representation with the TagProp method for image auto-annotation, and the results shows that using the Flickr-tags in combination with visual features improves the results of any method using only visual features. The university of Amsterdam’s concept detection system [13] focused on the per-image evaluation by modifying the probabilistic output of the SVM, they had disabled Platts conversion method to probabilities and used the distance to the decision boundary. Meiji University [11] devised a system that combines conceptual fuzzy sets with visual words, meanwhile constructed a system using Flickr User Tags. The Wroclaw University proposed a system [15] based on robustness of the global color and texture image features in connection with different similarity measures.



`{0A432C9F-1732-45E6-90F7-A6A7B75FA889}.jpg`

Flickr user tags (10) : peacock, bird, beautiful, pretty, feathers, waimea, waimeafalls, explore, animal, interestingness,

Fig. 1. A example image with Flickr user tags, which contains textual semantic concepts information: ’bird’, ’beautiful’, ’interesting’, etc.

The state of the art works above rely on visual features (including color, texture, shape, high level, and SIFT)[3, 4] or textual features, and visual features have generally been elaborated to catch the visual semantic details or atmosphere in images [5, 8]. However, these approaches may fail when a semantic concepts is not clear or obvious in images. This is the case especially for some abstract and emotional concepts which are subjective and cannot well defined by visual factors. Therefore additional information is needed, and we propose in this paper to make use of textual information describing the image which is provided by Flickr user tags, Figure 1 illustrates the additional information as provided by the Flickr user tag ('bird', etc.) with respect to semantic concepts (animal, bird, etc.). Thus, we developed two textual features to catch the semantic meanings of image tags. The framework is illustrated in Figure 2.

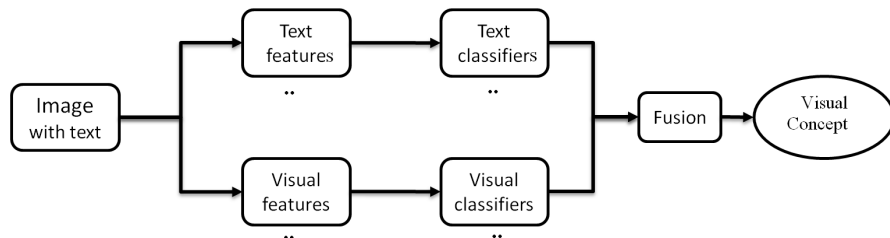


Fig. 2. The framework of our approach. For each image, the associated text is summarized to build the text features for text classifiers. Meanwhile, visual descriptors produce visual features for visual classifiers. Classifiers are combined to predict the semantic concept of the input image.

The rest of this paper is organized as follows. Section 2 presents 10 text-based features and 24 visual features representing emotion semantics. Section 3 provides experiments and results. Finally, section 4 draws the conclusion.

2 Features for semantic concepts

In this section, we firstly present two text-based features representing semantic concepts [24](Section 2.1), following with (Section 2.2) description of visual features which can be categorized into four groups: color, texture, shape and high level[23].

2.1 Textual features

Recently, Wang gang et al. [14] build text image feature for image object classification and demonstrate that it improves performance of visual object classification based on words frequency on a large similar image database. In contrast to this work, our approach is to build text features for concept detection based on a semantic distance between words, which are expected to capture the

semantic meanings from images and more directly reflects the semantics of the scene in images.

Method 1(textM1): The basic idea is to calculate semantic distance between the text surrounding images and an dictionary based on path similarity, denoting how similar two word senses are based on the shortest path that connects the senses in taxonomy. Firstly, we build a dictionary by using 99 concepts category name, which consists of 119 single words. After preprocessing on stop-words, we calculate the semantic distance matrix between the text associated image and dictionary based on a WordsNet by using a Natural Language Toolkit [16]. At last, the semantic distance feature is build based on the words semantic distance matrix. The procedure of Method 1 is as following:

Method 1: Procedure of text method 1 is based on path distance. The semantic dictionary $D, D = d$.

Input: labels data W and dictionary $D = \{d_i\}$

Output: text feature; $|f| = d, 0 < f_i < 1$.

- preprocess the tags by using a stop-words filter.
 - if image has no tags $W = 0$, return $f, f_i = 1/2$
 - Do for each words $w_t \in W$:
 1. if the path distance of w_t and d_i cannot be found, set $S(t, i) = 0$.
 2. Calculate the path distance $dist(w_t, d_i)$, where $dist$ is a simple node counting in the path from w_t to d_i .
 3. Calculate the path similarity as: $S(t, i) = 1/(dist(w_t, d_i) + 1)$,
 - Calculate the feature f as: $f_i = \sum_t S(t, i)$, and normalize it to $[0 1]$.
-

Method 2 (textM2): The idea of is to directly measure the emotional ratings of valence and arousal dimensions by using the Affective Norms for English Words (ANEW) set [17], which is being developed to provide a set of normative emotional ratings (including valence, arousal dimension) for a large number of words in English language. The procedure for Method 2 is as following:

Method 2: Procedure of textM2 is by using ANEW database, which contains dictionary D . and ratings of valence V and arousal A of each words in $D, |D| = d$.

Input: labels data W , dictionary D ratings of valence V and arousal A .

Output: text feature; $|f| = 2, 0 < f_i < 9$.

- preprocess the tags by using a stop-words filter.
 - if image has no tags $W = 0$, return $f, f_i = 5$
 - Do for each words $w_t \in W$:
 1. if the path distance of w_t and d_i cannot be found, set $S(t, i) = 0$,
 2. Calculate the path distance $dist(w_t, d_i)$, where $dist$ is a simple node counting in the path from w_t to d_i ,
 3. Calculate the path similarity as: $S(t, i) = 1/(dist(w_t, d_i) + 1)$,
 - Calculate the distance vector $m_i = \sum_t S(t, i)$, and normalize it to $[0 1]$,
 - Calculate the feature f as: $f_1 = (1/d) \sum_i (m_i.V_i)$, and $f_2 = (1/d) \sum_i (m_i.A_i)$,
-

We build 10 textual features based on above two methods on different words semantic distance(path and wup) and dictionary(dict119 and dict1034). Mean-

while, we modified textM1 and textM2 using different calculation methods including sum and max. A summary of text features illustrated in Table 1.

Table 1. The summary of textual features.

No.	Short name	Description
1	textM1_ftr1	obtained by using path distance on 119 words dictionary
2	textM1_ftr2	obtained by modifying the sum calculation to max method, and use path distance on 119 words dictionary
3	textM1_ftr3	obtained by using wup distance on 119 words dictionary
4	textM1_ftr4	obtained by modifying the sum calculation to max method, and use wup distance on 119 words dictionary
5	textM1_ftr5	obtained by using path distance on 1034 words dictionary
6	textM1_ftr6	obtained by modifying the sum calculation to max method, and use path distance on 1034 words dictionary
7	textM1_ftr7	obtained by using wup distance on 1034 words dictionary
8	textM1_ftr8	obtained by modifying the sum calculation to max method, and use wup distance on 1034 words dictionary
9	textM2_ftr9	obtained by modifying the sum calculation to max method, and use path distance on 1034 words dictionary
10	textM2_ftr10	obtained by using path distance on 1034 words dictionary

2.2 Visual features

We introduce various visual features to describe interesting details or to catch the global image atmosphere representing semantic concepts including following 4 groups, showing in Table 2:

Table 2: The summary of visual features.

No.	Short name	Description
11	color_hist	Histogram is concatenated from 128 bins on gray image level.
12	color_hsv	histogram is concatenated from 64 bins on each HSV channel.
13	color_moment	three central moments(Mean, Standard deviation and Skewness) on HSV channels.
14	color_msb	mean saturation and brightness
15	color_pad	approx. emotional coordinates based on brightness and saturation according to Valdez and Mehrabian [26]
16	texture_lbp	a compact multi-scale texture descriptor dealing with various changes in lighting and viewing condition

17	texture_hsvlbp	four multi-scale color LBP operators in order to increase photometric invariance property and discriminative power of the original LBP operator, according to Zhu’s work [28]
18	texture_invlbp	
19	texture_rgblbp	
20	texture_oppolbp	
21	texture_tamura	features by Tamura [18] including coarseness, contrast, directionality.
22	texture_cooccu	described by Haralick (1973), defined over an image to be the distribution of co-occurring values at a given offset.
23	texture_autocorr	referred to as the autocorrelation coefficient, which is a mathematical tool for finding repeating patterns.
24	shape_histLine	12 different orientations by using Hough transform.
25	highlevel_harm	try to describe color harmony of images based on Itten’s color theory [9]
26	hlevel_dyn	oblique lines communicate dynamism and action whereas horizontal or vertical lines rather communicate calmness and relaxation.
27	hlevel_aestheticDatta	implement most of the features (44 of 56) except those that are related to IRM (integrated region matching) technique [19].
28	hlevel_aestheticYke	implement Y.Ke’s aesthetic criteria including: spatial distribution of edges, hue count, blur, contrast and brightness [20].
29	hlevel_facect	implement the face counting method according to ViolaJones’s work. [29]
30	siftfeature_c	uses the C invariant, and scale-invariant with respect to light intensity
31	siftfeature_rgb	computed for each RGB channel independently.
32	siftfeature_hsv	computed for each HSV channel independently.
33	siftfeature_oppo	describes all the channels in the opponent color space using SIFT features.
34	siftfeature_daisy	based on daisy descriptors on SIFT interest point

Note that all SIFT features are computed using bag-of-words modelling with 4000 codewords and hard assignment.

3 Experiments and Results

3.1 Submitted runs

All runs are based on above descriptors including 10 textual ones and 24 visual ones, and that we do not use the EXIF meta data provided for the photos. During the experiments, we spot that Libsvm [25] performs better than

Adaboost classifier, not only on mean average precision but also the training speed. Thus, all runs are obtained by LIBSVM classifiers based on two kinds of kernels Chi-square kernels (Feature 11,12,16,17,18,19,20,30,31,32,33,34) or RBF kernels (other features). In order to obtain a stable and better performance, Firstly, we divided the training set into train part (50%, 4005 images) and validation part(50%, 3995 images), and we conducted experiments to select best weight to balance the positive and negative samples, then we obtained mean average precision for each feature. We performed our runs based on following configuration:

1. **textual model** we selected top 4 features among 10 textual features for each concept according to mAP, and use the mAP as score to combine the output of probability measurements of classifiers. We selected the threshold based on distribution of the training set.
2. **textual + visual model** we selected top 21 features among 34 visual and textual features for each concept according to mAP, and use the mAP as score to combine the output of probability measurements of classifiers. We selected the threshold based on best F-measure on validation set.
3. **textual model** we selected top 5 features among 10 textual features for each concept according to mAP, and use the mAP as score to combine the output of probability measurements of classifiers. We selected the threshold based on best F-measure on validation set.
4. **visual model** we selected top 5 features among 24 visual features for each concept according to mAP, and use the mAP as score to combine the output of probability measurements of classifiers. We selected the threshold based on best F-measure on validation set.
5. **textual + visual model** we selected top 22 features among 24 visual and textual features for each concept according to mAP, and use the mAP as score to combine the output of probability measurements of classifiers. We selected the threshold based on distribution of the training set.

Note that during the experiments, we perform fusion methods including min, max, mean, score(mAP as the score), and selected best fusion among 4 methods(min, max, mean, score) for each concept, and we find that score fusion method is more stable and better, especially when we carried out the same experiments on PASCAL2007 by using visual features only. Thus, we chose score one for all runs.

3.2 Results

We submitted 5 runs based on above configuration and features, and among the 5 runs, the 5th one achieved the best performance, which indicated that the combination of textural and visual features outperform than the other runs. We find that combination of 22 features received better results than the 21 ones. Moreover, we conducted experiments on validation set, and found that if the number of combining features is properly selected, it can improve 5 to 10

% points on mAP performance. Thanks to combination of textual and visual features, our 5th textual and visual model ranked 2^{end} out of 79 runs.

Table 3. The results of our submitted runs.

Submitted runs	mAP(%)	F-ex(%)	SR-Precision(%)
text_model_1	31.76	43.17	67.49
visual_text_model_2	42.96	57.57	71.74
text_model_3	32.12	40.97	67.57
visual_model_4	35.54	53.94	72.50
visual_text_model_5	43.69	56.69	71.82

4 Conclusion

This is the first year we take part into ImageCLEF2011, and we focus on the features and expect that the textual feature can be used to improve the performance of visual features. The experiments show that our textual features combining visual ones outperform the best visual ones. Our best textual and visual model obtains 43.69% in mAP, about 9% higher than the best visual-only model.

Besides we have conducted experiments to compare different classifiers and fusion methods, and we found that SVM classifier performed better than the Adaboost one, and the score fusion method provided a more stable and better results.

5 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. Flickr, <http://www.flickr.com>
2. Gettyimages, <http://www.gettyimages.com>
3. A. W. M Smeulders, et al., Content-based Image Retrieval: the end of the early years. IEEE Trans. PAMI, vol. 22, no.12, pp.1349-1380, (2000).
4. Z. Zeng et al. A survey of affect recognition methods: audio, visual and spontaneous expressions. IEEE Transactions PAMI, 31(1):39-58, (2009).
5. W. Wang, Q. He. A survey on emotional semantic image retrieval. ICIP, pp. 117-120, (2008).
6. Huiskes,M.,Lew,M.:The MIR Flickr retrieval evaluation.In:ACM MIR(2008).
7. Nowak,S.,Huiskes,M.:New strategies for image annotation: Overview of the photo annotation task at imageclef 2010.In:In the Working Notes of CLEF 2010. (2010).

8. C. Columbo, A. Del Bimbo, P. Pala. Semantics in visual information retrieval. *IEEE Multimedia*. 6(3):38-53, (1999).
9. J. Itten. *The art of colour*. Otto Maier Verlag, Ravensburg, Germany, (1961).
10. E Dellandrea, N. Liu, L. Chen, Classification of affective semantics in images based on discrete and dimensional models of emotions. *International Workshop on CBMI*. pp. 99-104.(2010).
11. Naoki Motohashi, Ryo Izawa, and Tomohiro Takagi: Meiji University at the ImageCLEF2010 Visual Concept Detection and Annotation Task: Working notes In:In the Working Notes of CLEF 2010. (2010).
12. Thomas Mensink¹, Gabriela Csurka, Florent Perronnin, Jorge Sanchez, and Jakob Verbeek: LEAR and XRCEs participation to Visual Concept Detection Task - ImageCLEF 2010.In:In the Working Notes of CLEF 2010. (2010).
13. Koen E. A. van de Sande and Theo Gevers :The University of Amsterdam's Concept Detection System at ImageCLEF 2010.In:In the Working Notes of CLEF 2010. (2010).
14. Gang Wang, Derek Hoiem, and David Forsyth, Building text features for object image classification. *CVPR*, 1367 - 1374, 20-25 June (2009).
15. Michal Stanek, Oskar Maier, and Halina Kwasnicka.:The Wroclaw University of Technology Participation at ImageCLEF 2010 Photo Annotation Track.In:In the Working Notes of CLEF 2010. (2010).
16. Natural language toolkit. <http://www.nltk.org>
17. Bradley M.M., Lang P.J., Affective norms for English words (ANEW). Tech. Rep C-1, GCR in Psychophysiology, University of Florida, (1999).
18. H. Tamura, S. Mori, and T. Yamawaki. Textural features corresponding to visual perception. *IEEE Transactions on SMC*, 8(6):460-473, June (1978).
19. R. Datta, J. Li, James Z. Wang, Content-based image retrieval: approaches and trends of the new age, *ACM Workshop MIR*, Nov. 11-12, Singapore, (2005).
20. Y. Ke, X. Tang, and F. Jing. The Design of High-Level Features for Photo Quality Assessment, *Proc. CVPR*, (2006).
21. P. Dunker, S. Nowak, A. Begau, C. Lanz. Content-based mood classification for photos and music. *ACM MIR*, pp. 97-104, (2008).
22. Lang P.J., Bradley M. M., B. N. Cuthbert, the IAPS: Technical manual and affective ratings. Tech. Rep A-8., GCR in Psychophysiology, Univ. of Florida, (2008).
23. N. Liu, E Dellandréa, B. Tellez, L. Chen.: Evaluation of Features and Combination Approaches for the Classification of Emotional Semantics in Images. *IC. on VISAPP*, Mar 5-7, Portugal, (2011).
24. N. Liu, E Dellandréa, B. Tellez, L. Chen.: Associating Textual Features with Visual Ones to Improve Affective Image Classification. *IC. on ACII*, Oct 13-17, Tennessee, USA (2011).
25. Chih-Chung Chang and Chih-Jen Lin, LIBSVM : a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1– 27:27, 2011. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>. (2011).
26. P. Valdez and A. Mehrabian. Effects of color on emotions. *Journal of Experimental Psychology: General*, 123(4):394-409, 1994. (1994).
27. J. Machajdik and A. Hanbury.: Affective Image Classification using Features Inspired by Psychology and Art Theory, *ACM Multimedia 2010 - Multimedia Content Track Full Paper*, Florence, Italy.(2010).
28. Chao Zhu, Charles-Edmond Bichot, Liming Chen.: Multi-scale Color Local Binary Patterns for Visual Object Classes Recognition, 2010 *I.C in ICPR*. (2010).
29. Paul Viola, Micheal J. Jones.: *International Journal of Computer Vision* 57(2), 137154, 2004. (2004).