Ensemble Approach Based on Conditional Random Field for Multi-Label Image and Video Annotation*

Xin-Shun Xu[1,2]
xuxinshun@sdu.edu.cn

Yuan Jiang[2]
jiangy@lamda.nju.edu.cn

Liang Peng[1]
pengliang.sdu@gmail.com

Xiangyang Xue[3]
xyxue@fudan.edu.cn

Zhi-Hua Zhou[2]
zhouzh@lamda.nju.edu.cn

[1]School of Computer Science and Technology, Shandong University, Jinan 250101, China
[2]National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093, China
[3]School of Computer Science, Fudan University, Shanghai 200433, China

ABSTRACT

Multi-label image/video annotation is a challenging task that allows to correlate more than one high-level semantic keyword with an image/video-clip. Previously, a single model is usually used for the annotation task, with relatively large variance in performance. The correlation among the annotation keywords should also be considered. In this paper, to reduce the performance variance and exploit the correlation between keywords, we propose the En-CRF (Ensemble based on Conditional Random Field) method. In this method, multiple models are first trained for each keyword, then the predictions of these models and the correlations between keywords are incorporated into a conditional random field. Experimental results on benchmark data set, including Corel5k and TRECVID 2005, show that the En-CRF method is superior or highly competitive to several state-of-the-art methods.

Categories and Subject Descriptors

I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding—*video analysis*; I.2.6 [**Artificial Intelligence**]: Learning

General Terms

Algorithms, Design, Experimentation

Keywords

Image Annotation, Video Annotation, Conditional Random Field, Multi-Label Learning, Ensemble Methods

1. INTRODUCTION

With the fast accumulation of image and video data, how to effectively access large-scale multimedia database, such as indexing, browsing and retrieval, has become a challenging task. As most users prefer textual queries, a common theme is to first annotate these images/videos with keywords which describe the content in the images, and then use these keywords for indexing or browsing. Manual annotation is labor-intensive and time-consuming. Thus, automatic image/video annotation has emerged as an important topic, and become an active research topic in recent years. This paper focuses on multi-label image/video annotation task in which an image/video-clip is associated with more than one keyword simultaneously. The learning framework to tackle such scenario is called multi-label learning in machine learning community which has attracted more and more attentions during the past few years [21, 30, 31, 23].

Many techniques have been developed for the image/video annotation task. For image annotation, Jeon et al. [8] proposed the cross-media relevance model (CMRM) which tries to estimate the joint probability of the visual keywords and the annotation keywords on the training data set. This relevance model was further improved through continuous-space relevance model (CRM) [11], multiple Bernoulli relevance model (MBRM) [4], and dual cross-media relevance model [13]. Carneiro et al. [3] proposed a probabilistic approach for this task. Guillaumin et al. [7] proposed a discriminatively trained nearest neighbor model in which tags of test images are predicted using a weighted nearest-neighbor model to exploit labeled training images. In [32], Zhang et al. introduced a regularization based feature selection algorithm to leverage both the sparsity and clustering properties of features, and incorporated it into the image annotation task.

To consider multiple labels, a common approach is to learn a separate classifier for each keyword on training data, and use them to predict whether a new image belongs to some classes. In this paradigm, SVM based methods have been shown effective [1, 26]. Gao et al. proposed a hierarchical boosting algorithm by incorporating feature hierarchy and boosting to scale up SVM image classifier. Yanagawa et al. [25] built a set of baseline detectors for 374 LSCOM concepts by using SVM. Those methods also treat the concepts independently. In most image/video annotation tasks, there are significant correlations in keywords. To exploit the correlation, and improve the performance of the binary concept classifiers, some methods try to construct models with a context-based concept fusion strategy [9, 14, 18, 22, 28]. Qi et al. [20] simultaneously modeled both the individuals as well as their correlations with a unified formulation. Recently, a new machine learning framework named MIML (Multi-Instance Multi-Label learning) was proposed [34], which represented each object as multiple

*Area Chair: Kiyoharu Aizawa. This research was supported by NSFC (60975043,61173068), CPSF (20100470063), Jiangsu PPPRF (1001001B), IIFSDU (2009TS033), and NSF of Shandong Province (Q2008G06).

MM'11, November 28–December 1, 2011, Scottsdale, Arizona, USA.
Copyright 2011 ACM 978-1-4503-0616-4/11/11 ...$10.00.

instances and allowed the object to have multiple labels simultaneously. This framework was soon applied to and found well useful in image annotation [27].

Ensemble methods are a kind of powerful learning paradigm which is able to achieve strong generalization ability by using multiple learners [33]. Usually, ensemble methods combine the individual learners by voting, averaging, or their variants. In this paper, we propose the En-CRF method, which combines multiple models, as potential functions, with a conditional random field [10]. In addition to the profit from combining multiple models, the En-CRF method also takes benefit from using CRF to exploit the correlations between the keywords.

The rest of this paper is organized as follows: In section 2, we introduce the En-CRF method, including the training and testing processes. In section 3, we report on the experimental results on benchmark data set Corel5k and TRECVID 2005. Finally, we conclude this paper in section 4.

2. THE EN-CRF METHOD

2.1 The Framework

Given a feature vector X, let $Y \in \mathcal{Y}$ denote the label vector with length m, and y_i is the ith element of Y, taking the value in {-1, +1} with +1 denotes that the corresponding label is a proper one for X, while -1 denotes that the corresponding label is irrelevant to X. For each keyword, we construct r classifiers. These classifiers can be different models, e.g., SVM with different kernels, neural networks with different number of hidden units, etc. The posterior probability for label Y is of the following form:

$$p(Y|X) = \frac{1}{Z(X)} \exp\left(-E(Y|X)\right) \tag{1}$$

where $E(Y|X)$ is the energy function defined as:

$$E(Y|X) = \sum_{i}^{r}\sum_{j}^{m} w_{ij} f_{ij}(X, y_j) + \sum_{i}^{m}\sum_{j}^{m}\sum_{k=0}^{3} w_{ijk}\delta(y_i, y_j, k) f'_{ij}(y_i, y_j) \tag{2}$$

where $f_{ij}(X, y_j)$ is defined to measure how likely the result of the ith classifier for the jth label is positive. For instance, as used in this paper, its value can be the signed distance to the hyperplane when SVM is used as the classifier. $f'_{ij}(y_i, y_j)$ measures the correlation between the ith and jth keywords, $k \in \{0, 1, 2, 3\}$ corresponds to four combinatorial states of two labels, e.g., 0 corresponding to the state that both labels are negative, and $\delta(y_i, y_j, k)$ is defined as:

$$\delta(y_i, y_j, k) = \begin{cases} 1 & y_i = -1 \text{ and } y_j = -1 \text{ and } k = 0 \\ 1 & y_i = -1 \text{ and } y_j = +1 \text{ and } k = 1 \\ 1 & y_i = +1 \text{ and } y_j = -1 \text{ and } k = 2 \\ 1 & y_i = +1 \text{ and } y_j = +1 \text{ and } k = 3 \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

Similarly, $Z(X)$, the partition function, is defined by:

$$Z(x) = \sum_{Y\in\mathcal{Y}} \exp\left(-E(X|Y)\right) \tag{4}$$

2.2 Training Process

The training process of En-CRF consists of two steps: (1) setting up $f_{ij}(X, y_j)$ and $f'_{ij}(y_i, y_j)$; (2) learning the parameters w.

As mentioned above, $f_{ij}(X, y_i)$ is defined to measure how likely the result of the ith classifier for the jth label is positive. For this, we first train r SVM's with different kernels for each class, and therefore get r hyperplanes for each class. Then, $f_{ij}(X, y_j)$ takes the value of the signed distance of X to the ith hyperplane. $f'_{ij}(y_i, y_j)$ measures the correlation of the ith and jth keywords, and it can be realized by the co-occurrence of the ith and jth keywords. Notice that some keyword pairs may have rather weak interactions. That is, the presence/absence of one keyword would not affect the presence/absence of another keyword. Thus, it is not necessary to involve all classes during training process. Here, we adopt the normalized mutual information [20] to determine which label pairs should be selected. For keywords p and q, the normalized mutual information is defined as

$$NormMI(p, q) = \frac{MI(p, q)}{min\{H(p), H(q)\}} \tag{5}$$

where $MI(p, q)$ is the mutual information of words p and q, and $H(p)$ is the marginal entropy of word p as defined respectively,

$$MI(p, q) = \sum\nolimits_{y_p, y_q} P(y_p, y_q) log \frac{P(y_p, y_q)}{P(y_p)P(y_q)} \tag{6}$$

$$H(p) = -\sum\nolimits_{y_p\in\{0,1\}} P(y_p)P(y_q) \tag{7}$$

Based on the normalized mutual information, the pairs whose correlations are larger than a threshold are considered.

Once we get $f_{ij}(X, y_j)$ and $f'_{ij}(y_i, y_j)$, given the training set $D = \{(X^1, Y^1), (X^2, Y^2), ..., (X^n, Y^n)\}$, the penalized log likelihood of parameters w is

$$L(w|D) = \log\left(\prod_{i=1}^{n} p(Y^i|X^i)\right) - \sum_{i,j} \frac{w_{ij}^2}{2\sigma^2} - \sum_{i,j,k} \frac{w_{ijk}^2}{2\sigma^2} \tag{8}$$

where the last two terms are introduced to reduce overfitting. Given the gradient of the log likelihood at w_{ij} or w_{ijk}, optimization methods such as BFGS [2] can be used to find the optimal w.

2.3 Testing and Ranking Process

A trained En-CRF model can be easily used to annotate a new example X. There are also two steps to annotate a new example. First, we should calculate $f_{ij}(X, y_j)$, and δ used in Eq. 2 (note that f'_{ij} can be computed in advance). $f_{ij}(X, y_j)$ can be easily computed based on the hyperplane of ith classifier for the jth label. The second step is to infer a label vector Y for this X according to

$$Y = \underset{Y\in\mathcal{Y}}{argmax}\, p(Y|X). \tag{9}$$

This inference problem can be executed efficiently by some dynamic programming algorithms, e.g., the Viterbi-like algorithm used in this paper.

It would be nice if the annotated keywords can be ranked according to their relevancy to the example. This requires us to give a score to every annotated keyword. Here we can get this by computing the conditional expectation [20] of a keyword. Suppose that the predicted label vector for an image is Y, and $\wp$ is the keyword set, the conditional expectation of y_p for keyword p can be computed by

$$\mathcal{E}(y_p|X, Y_{\wp\backslash p}) = P(y_p = +1|X, Y_{\wp\backslash p}) - P(y_p = -1|X, Y_{\wp\backslash p}) \tag{10}$$

According to Eq. 1, $P(y_p|X, Y_{\wp\backslash p})$ is defined as

$$P(y_p|X, Y_{\wp\backslash p}) = \frac{1}{Z_p} \exp\left(-E(y_p \circ Y_{\wp\backslash p}|X, w)\right) \tag{11}$$

where

$$Z_p = \sum\nolimits_{y_p\in\{+1,-1\}} \exp\left(-E(y_p \circ Y_{\wp\backslash p}|X, w)\right) \tag{12}$$

and the circle operator means to concatenate two parts of labels into one.

The En-CRF method is summarized in Algorithm 1.

Algorithm 1 En-CRF

```
Training
1. Input: The training data set D
2. Train $r$ SVM's for each class label, and get the hyperplanes $H$.
3. Compute $f_{ij}$, $f'_{jl}$ and $\delta$ for all data in D
4. Get $w$ by optimizing Eq. 8.
Testing
1. Input: a test example X
2. Compute $f_{ij}$ according to the hyperplanes $H$, and $\delta$.
3. Infer the label vector $Y = argmax_{Y \in \mathcal{Y}}\, p(Y|X)$.
4. Output: Y
```

3. EXPERIMENTS

3.1 Experiments on Image Data

In this section, En-CRF is tested on the Corel data set which contains 5,000 images collected from larger Corel CD set. Each image has been annotated by an average of 3–5 keywords from a dictionary of 371 keywords. The data set is further divided into two parts: training set and testing set which contains 4500 and 500 images, respectively. Finally, the dictionary contains 260 words that appear in both the training and testing set.

Similar to the routine in [24], images are first represented with pyramid structure in which the division scheme is 1×1, 2×2, 1×3. From each region, SIFT, C-SIFT, rgSIFT, and OpponentSIFT descriptors are first extracted with both Harris-Laplace point sampling and dense sampling methods. Then, k-means is used to generate a dictionary with size 2000. Finally, for each type of feature, a region is represented as a 'bag-of-words' histogram.

For each keyword, 6 SVM's with different kernels are trained on the training set. These kernels include (1)Spatial pyramid kernel with weighted χ^2 distances of multiple features [24]; (2)Spatial pyramid kernel with weighted earth mover distance (EMD) [29] of multiple features, in which EMD is used instead of χ^2 distance; (3)Pyramid match kernel [6]; (4)Multiple instance kernel [5], in which each feature is treated as an instance; (5) RBF kernel, in which all features are sequentially combined into a flat vector. (6)Polynomial kernel, in which all features are also combined into one vector as in (5). Note that the features used in (2)&(3) is a little different from those used in other kernels. For saving space, here we do not give the detail.

The results are shown in Table 1. The left part is the annotation results in which 'P' is mean precision, 'R' is mean recall and 'N+' denotes the number of keywords with non-zero recall value. The right part is the retrieval results in which the Mean Average Precision (MAP) is used to measure the performance. From Table 1, we can find that En-CRF is superior or highly competitive to state-of-the-art methods.

3.2 Experiments on Video Data

In this section, we evaluate En-CRF on the video benchmark data set TRECVID2005. In the experiments, only the development set is used which contains 80 hours international broadcast news in Arabic, English and Chinese. These news videos are first automatically segmented into 61901 subshots [19]. For each subshot, a keyframe is extracted to represent the subshot, and associated with one or more concepts of 39 concepts according to LSCOM-Lite annotations [17]. Most videos are multi-labeled, e.g., 75.7% shots/subshots have more than one concept. The feature extraction scheme and kernels are as same as those used in subsection 3.1. The performance is evaluated by Average Precision (AP) and MAP which are the official performance metrics in TRECVID evaluations. The AP corresponds to the area under a non-interpolated recall/precision curve and it favors highly ranked relevant subshots. In this paper, the AP is calculated on the top-ranked 2000 shots returned from these methods. The MAP are calculated by averaging the AP's of all concepts. We compare En-CRF with CML (II) and CMLT [20], both are popular methods which considering correlations among concepts, where CMLT incorporates temporal information into a kernel. All parameters are selected by 3-fold cross-validation on training set. The AP and MAP are shown in Figure 1. From Figure 1, we can find that

- En-CRF achieves the best performance on 28 concepts, and performs better on 32 concepts than CML(II), and 28 than CMLT. En-CRF obtains the best MAP.
- En-CRF performs much better than CML(II) and CMLT on some concepts, e.g., "Office"(82.34% better than CML(II), 65.37% better than CMLT), and "Urban"(68.57% better than CML(II), and 52.88% better than CMLT).
- En-CRF deteriorates on some concepts, e.g., "Sports" due to the fact the there are strong temporal consistency in video data for such concepts, and such information is not considered in En-CRF.

Table 1: Comparison of En-CRF and state-of-the-art methods on Corel5K. 'P' denotes mean precision, 'R' denotes mean recall, 'N' denotes the number of keynotes with non-zero recall, and 'MAP' means mean average precision

Annotation				Retrieval (MAP)	
Method	P	R	N+	260 words	recall≥ 0
infNet[16]	0.17	0.24	112		
JEC[15]	0.27	0.32	139		
LASSO[15]	0.24	0.29	127		
TagProp[7]	0.33	0.42	160		
GS[32]	0.30	0.33	146		
CRM[11]	0.16	0.19	107	0.26	0.30
SML[3]	0.23	0.29	137	0.31	0.49
TGLM[12]	0.25	0.29	131	0.29	0.52
MBRM[4]	0.24	0.25	122	0.30	0.35
En-CRF	0.32	0.33	148	0.32	0.42

4. CONCLUSIONS

In this paper, we propose the En-CRF method for image and video annotation. In En-CRF, multiple classifiers are trained for each keyword, and all classifiers are combined with a conditional random field. Experimental results on benchmark data sets, including Corel5k and TRECVID 2005, show that En-CRF is superior or highly competitive to several state-of-the-art methods. An interesting future work is to improve En-CRF by incorporating helpful temporal and structured information in video/image annotation tasks.

5. REFERENCES

[1] A. Amir, M. Berg, and S.-F. Chang et al. IBM research TRECVID-2003 video retrieval system. In *TRECVID*, 2003.

[2] R. H. Byrd, J. Nocedal, and R. B. Schnabel. Representations of quasi-newton matrices and their use in limited memory methods. *Mathematical Programming*, 63(2):129–156, 1994.

[3] G. Carneiro, A. B. Chan, P. J. Moreno, and N. Vasconcelos. Supervised learning of semantic classes for image annotation and retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(3):394–410, 2007.

[4] S. Feng, R. Manmatha, and V. Lavrenko. Multiple bernoulli relevance models for image and video annotation. In *CVPR*, 1002–1009, 2004.

[5] T. Gartner, P. A. Flach, A. Kowalczyk, and A. J. Smola. Multi-instance kernels. In *ICML*, 179–186, 2002.

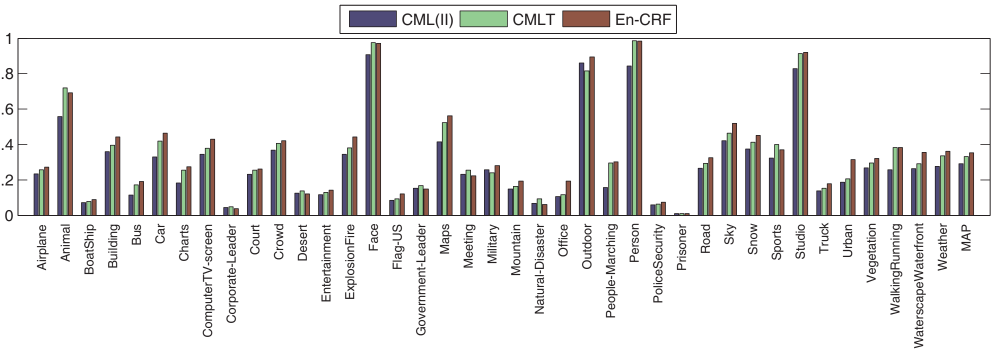

Figure 1: The performance of CML(II), CMLT and En-CRF

[6] K. Grauman and T. Darrell. The pyramid match kernel: Discriminative classification with sets of image features. In *ICCV*, 1458–1465, 2005.

[7] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid. Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In *ICCV*, 309–316, 2009.

[8] J. Jeon, V. Lavrenko, and R. Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In *SIGIR*, 119–126, 2003.

[9] W. Jiang, S.-F. Chang, and A. C. Loui. Context-based concept fusion with boosted conditional random fields. In *ICASSP*, 949–952, 2007.

[10] J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, 282–289, 2001.

[11] V. Lavrenko, R. Manmatha, and J. Jeon. A model for learning the semantics of pictures. In *NIPS 16*, 553–560, 2003.

[12] J. Liu, M. Li, Q. Liu, H. Lu, and S. Ma. Image annotation via graph learning. *Pattern Recognition*, 42(2):218–228, 2009.

[13] J. Liu, B. Wang, M. Li, Z. Li, W.-Y. Ma, H. Lu, and S. Ma. Dual cross-media relevance model for image annotation. In *MM*, 605–614, 2007.

[14] Z. Lu, H. H. Ip, and Q. He. Context-based multi-label image annotation. In *CIVR*, 279–287, 2009.

[15] A. Makadia, V. Pavlovic, and S. Kumar. A new baseline for image annotation. In *ECCV*, 316–329, 2008.

[16] D. Metzler and R. Manmatha. An inference network approach to image retrieval. In *CIVR*, 42–50, 2004.

[17] M. Naphade, L. Kennedy, J. R. Kender, S. F. Chang, P. Over, and A. Hauptmann. A light scale concept ontology for multimedia understanding for TRECVID 2005. In IBM Research Technical Report, 2005.

[18] M. R. Naphade, I. Kozintsev, and T. Huang. Factor graph framework for semantic video indexing. *IEEE Transactions on Circuits and Systems for Video Technology*, 12(1):40–52, 2002.

[19] C. Petersohn. Fraunhofer HHI at TRECVID 2004: Shot boundary detection system. In *TRECVID*, 2004.

[20] G.-J. Qi, X.-S. Hua, Y. Rui, J. Tang, T. Mei, M. Wang, and H.-J. Zhang. Correlative multilabel video annotation with temporal kernels. *ACM Transactions on Multimedia Computing, Communications and Applications*, 5(1):1–27, 2008.

[21] R. E. Schapire and Y. Singer. Boostexter: A boosting-based system for text categorization. *Machine Learning*, 39(2-3):135–168, 2000.

[22] J. R. Smith, M. Naphade, and A. Natsev. Multimedia semantic indexing using model vectors. In *ICME*, 445–448, 2003.

[23] G. Tsoumakas and I. Katakis. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining*, 3(3):1–13, 2007.

[24] K. E. van de Sande, T. Gevers, and C. G. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1582–1596, 2010.

[25] A. Yanagawa, S.-F. Chang, L. Kennedy, and W. Hsu. Columbia universityąŕs baseline detectors for 374 lscom semantic visual concepts. Technical report, Columbia University ADVENT, 2007.

[26] J. Yuan, Z. Guo, and L. L. et al. THU and ICRC at TRECVID. In *TRECVID*, 2007.

[27] Z.-J. Zha, X.-S. Hua, T. Mei, J. Wang, G.-J. Qi, and Z. Wang. Joint multi-label multi-instance learning for image classification. In *CVPR*, 1–8, 2008.

[28] Z.-J. Zha, T. Mei, X.-S. Hua, G.-J. Qi, and Z. Wang. Refining video annotation by exploiting pairwise concurrent relation. In *MM*, 345–348, 2007.

[29] J. Zhang, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: a comprehensive study. *International Journal of Computer Vision*, 73(2):213–238, 2007.

[30] M.-L. Zhang and Z.-H. Zhou. Multilabel neural networks with applications to functional genomics and text categorization. *IEEE Transactions on Knowledge and Data Engineering*, 18(10):1338–1351, 2006.

[31] M.-L. Zhang and Z.-H. Zhou. ML-KNN: A lazy learning approach to multi-label learning. *Pattern Recognition*, 40(7):2038–2048, 2007.

[32] S. Zhang, J. Huang, Y. Huang, Y. Yu, H. Li, and D. Metaxas. Automatic image annotation using group sparsity. In *CVPR*, 3312–3319, 2010.

[33] Z.-H. Zhou. Ensemble learning. In S. Z. Li, editor, *Encyclopedia of Biometrics*, 270–273, Springer, 2009.

[34] Z.-H. Zhou and M.-L. Zhang. Multi-instance multi-label learning with application to scene classification. In *NIPS 19*, 1609–1616, 2007.