

Image Annotation Based on Multi-view Learning

Zhe Shi, Songhao Zhu^(✉), and Chengjian Sun

School of Automatic, Nanjing University of Posts and Telecommunications,
Nanjing 210046, China
njuptzsl@yeah.net

Abstract. With the explosive growth of image data collections on the web, labeling each image with appropriate semantic description based on the image content for image index and image retrieval has become an increasingly difficult and laborious task. To deal with this issue, we propose a novel multi-view semi-supervised learning scheme to improve the performance of image annotation by using multiple views of an image and leveraging the information contained in pseudo-labeled images. In the training process, labeled images are first adopted to train view-specific classifiers independently using uncorrelated and sufficient views, and each view-specific classifier is then iteratively re-trained using initial labeled samples and additional pseudo-labeled samples based on a measure of confidence. In the annotation process, each unlabeled image is assigned appropriate semantic annotations based on the maximum vote entropy principle and the relationship between annotations with respect to the results of each optimally trained view-specific classifier. Experimental results on a general-purpose image database demonstrate the effectiveness and efficiency of the proposed multi-view semi-supervised scheme.

1 Introduction

With the rapidly increasing volume of image data collection uploaded to the Internet, the demand of developing content-based analysis technologies to effectively organize, manage and utilize such huge amount of information resources has become an important and challenging research topic in the field of intelligent multimedia analysis. Among these technologies, image annotation, which aims to build an exact correspondence between visual information at the perceptual level and linguistic descriptions at the semantic level, is an elementary step and a promising step for content-based image indexing, retrieval and other related multimedia applications. Therefore, automatic annotating an image with high-level semantic descriptions has emerged as an important and challenging research topic in recent years.

In the past couple of years, many novel algorithmic techniques have been proposed to deal with the problem of image annotation, such as nested deep belief nets based method [1], covariance discriminative based method [2], bilinear deep learning based method [3], local and global information based method [4], separable principal components analysis based method [5], graph theory based method [6], two-dimensional

multi-label active learning based method [7], Wavelet feature based metric [8], high order statistics based method [9], sparse coding based method [10].

The task of these existing algorithms is to assign appropriate labels to a given image with respect to the semantic contents of the image. There are two issues that should be considered when designing an effective and efficient image annotation algorithm: on one hand, the number of labeled images is often very small while the number of unlabeled images is often very large; on the other hand, an image is generally represented by a combination of feature set, such as color, shape and texture. The performance of image annotation is seriously affected by the above discussed two issues. To address the first issue, the semi-supervised methods are adopted to leverage the information contained in unlabeled images to improve the prediction performance [11, 12]. To address the second issue, the multi-view learning algorithms are utilized to achieve the informative and representative training images to reduce the amount of labeled samples required for training [13]. In multi-view learning, multiple classifiers are trained separately using several different views extracted from the labeled samples, then these view-specific classifiers assign labels to pseudo-labeled samples and the disagreement among different view-specific classifiers is utilized to selected additional pseudo-labeled samples, finally new view-specific classifiers are re-trained using the initial labeled samples and the newly pseudo-labeled unlabeled samples to improve the overall annotation performance. The idea of multi-view learning and semi-supervised learning can be effectively integrated, such as the method proposed by Zhang in [14].

In this paper, we propose a new multi-view semi-supervised learning scheme for automatic image annotation. The basic idea of the proposed scheme, as shown in Fig. 1, is described as follows. Firstly, uncorrelated and sufficient views, such as color histogram, wavelet texture and edge direction histogram, are extracted to train V view-specific classifier $\{h_1, h_2, \dots, h_v, \dots, h_V\}$ independently, and the labels of each pseudo-labeled image over each view is achieved by using the learned view-specific classifier. Secondly, initial labeled samples and pairs of pseudo-labeled samples with high confidence are utilized to iteratively re-train view-specific classifiers to build better classifiers for improving the model performance. Finally, unlabeled images are assigned the category annotation using the maximum vote entropy principle based one optimally trained view-specific classifiers, and other annotations on the annotation list can be obtained based on the correlation between annotations. The experimental results show that by taking advantage of both multi-view and multi-view the proposed approach significantly outperforms state-of-the-art methods.

The rest of the paper is organized as follows. The proposed image annotation scheme is detailed in Sect. 2. The experimental results and some discussions of the proposed image annotation scheme are presented in Sect. 3. Finally, this paper is concluded in Sect. 4.

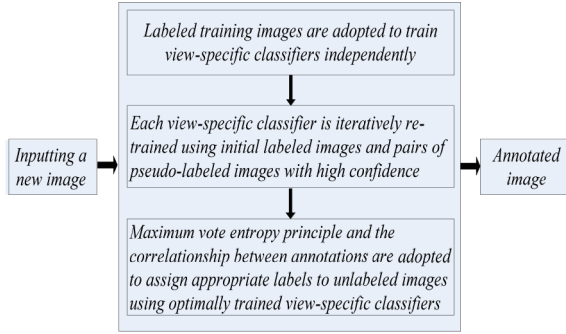


Fig. 1. The basic idea of the proposed scheme.

2 Multi-view Semi-supervised Annotation Scheme

In this section, we will detail the proposed multi-view semi-supervised learning scheme, which consists of multi-view classifiers learning process based on the uncorrelated and sufficient visual features, multi-view classifiers optimal process based on the initial labeled samples and pairs of pseudo-labeled samples with high confidence level, multi-view annotation process based on trained view-specific classifiers.

Before detailing the scheme of multi-view semi-supervised learning process, there are some notations that need to be pre-defined.

- An labeled image set LS with L images $\{x_1, x_2, \dots, x_l, \dots, x_L\}$ and a pseudo-labeled image set PS with U images $\{x_{L+1}, x_{L+2}, \dots, x_{L+u}, \dots, x_{L+U}\}$.
- An image x_l from labeled image set is attached with a set of annotations $Y = \{y_1, y_2, \dots, y_k, \dots, y_K\}$ where K is the number of annotations and each $y_k \in \{-1, 1\}$ indicates whether one semantic concept occurs (1) or not (-1).
- An image x_{L+u} from pseudo-labeled image set is represented as multi-view vector $\{x_{(L+u)1}, x_{(L+u)2}, \dots, x_{(L+u)v}, \dots, x_{(L+u)V}\}$ where V is the number of views.

2.1 Learning Process of Multi-view Classifiers

The process of each view-specific classifier learning is detailed as follows:

- Several distinct feature sets, such as color histogram, wavelet texture and edge direction histogram, are extracted and considered as uncorrelated and sufficient views of an image, then one view-specific classifier (Support Vector Machine) h_v on the v^{th} view is trained to predict decision boundaries between different categories using initial labeled image set $\{x_1, x_2, \dots, x_l, \dots, x_L\}$:

$$h_v : x_{lv} \rightarrow y_k, l \in (1, L), v \in (1, V), y_k \in Y \tag{1}$$

2.2 Semi-supervised Optimal Process of Multi-view Classifiers

The semi-supervised optimal process of multi-view classifiers is detailed as follows:

- Let p_{uv}^k denotes the probability of one pseudo-labeled sample $x_{(L+u)v}$ belongs to the k^{th} category in the v^{th} view:

$$p_{uv}^k = p_u^k(h_v) = p(y_k = 1 | x_{(L+u)v}) \quad (2)$$

The prediction of the pseudo-labeled sample $x_{(L+u)v}$ belonging to the k^{th} category in the v^{th} view can be computed using the following formula:

$$y_{uv}^k = \text{sgn}(p(y_k = 1 | x_{(L+u)v}) - p(y_k = -1 | x_{(L+u)v})) \quad (3)$$

For comprehensive evaluation of the confidence of each pseudo-labeled image's annotation results over all views, the Gaussian distance is here adopted to measure the disagreement level of annotation results over pseudo-labeled images using different view-specific classifier. Let $p_{uv} = \{p_{uv}^1, p_{uv}^2, \dots, p_{uv}^k, \dots, p_{uv}^K\}$ denotes the probability of the pseudo-labeled image $x_{(L+u)v}$ belongs to each category in the v^{th} view, and then a disagreement measurement between any two view-specific classifiers h_v and $h_{v'}$ over any a random pseudo-labeled image $x_{(L+u)v}$ can be formulated as follows:

$$\begin{aligned} DL(h_v, h_{v'}) &= \sum_{\{v, v'\} \subseteq V \wedge v \neq v'} \exp\left(-\|p_{uv} - p_{uv'}\|_2^2\right) \\ &= \sum_{\{v, v'\} \subseteq V \wedge v \neq v'} \exp\left(-\left(\sqrt{\sum_{k \in (1, K)} (p_u^k(h_v) - p_u^k(h_{v'}))^2}\right)^2\right) \end{aligned} \quad (4)$$

This notion of disagreement can then be extended to all view-specific classifiers by taking the average disagreement between any pair of view-specific classifiers:

$$\tilde{DL}(h_1, h_2, \dots, h_v, \dots, h_V) = \frac{2}{V(V-1)} \sum_{\{v, v'\} \subseteq V \wedge v \neq v'} DL(h_v, h_{v'}) \quad (5)$$

- According to the multi-view learning framework proposed by Sridharan et al. [15] for classification, a set of classifiers $\{h_{(v)}\}$ over each v view is optimally re-trained by maximizing their consensus on the unlabeled data. Specifically, each view-specific classifier is iteratively re-trained using initial labeled images and additional pseudo-labeled images, and the optimal procedure over each v view is repeated until the disagreement level of pseudo-labeled image set does not decrease after re-training. The optimal process of the multi-view semi-supervised classifier is detailed in Table 1.

Table 1. Optimal process of each multi-view semi-supervised classifier

Input:

- Training image collection containing labeled image set $LS(x_1, x_2, \dots, x_l, \dots, x_L)$ and pseudo-labeled image set $PS(x_{L+1}, x_{L+2}, \dots, x_{L+u}, \dots, x_{L+u})$;
- Size of the pair of two pseudo-labeled images S ;
- View-specific classifier set $\{h_1, h_2, \dots, h_v, \dots, h_V\}$;

Initialize:

- For each view v , train $h_v^k(0)$ using the labeled image set $LS(x_1, x_2, \dots, x_l, \dots, x_L)$;

Multi-View Semi-Supervised Classifier Learning:

- $t \leftarrow 0$;
- **repeat**
 - for** $s=1, \dots, S$ **do**
 - sample (u, u') from S
 - if** $p_u^k(h_v(t)) - p_{u'}^k(h_v(t)) \geq 0.25$ **then**
 - $LS \leftarrow LS \cup \{(x_{L+u}, +1), (x_{L+u'}, -1)\}$
 - else if** $p_{u'}^k(h_v(t)) - p_u^k(h_v(t)) < -0.25$ **then**
 - $LS \leftarrow LS \cup \{(x_{L+u}, -1), (x_{L+u'}, +1)\}$
 - end if**
 - end for**
 - $t \leftarrow t + 1$;
 - for each view v , train $h_v(t)$ over each category using the labeled image set LS ;
- **until** $\dot{DL}(h_1(t+1), \dots, h_V(t+1)) \geq \dot{DL}(h_1(t), \dots, h_V(t))$

output: $h_v(t), v \in \{1, 2, \dots, V\}$

During the iterative process of optimally training each view-specific classifier, the pairs of pseudo-labeled images with distinct difference in view-specific property are added into the initial labeled images to re-train each view-specific classifier to improve prediction performance of assigning appropriate labels for unlabeled images.

2.3 Annotation Process Using Multi-view Classifiers

For each image, there are many semantic annotations can be assigned to it based on the image content. Among these semantic annotations, the first annotation on the annotation list, also considered as the category annotation, should be first achieved since it semantically describes the main visual content. Other annotations on the annotation list can then be successively achieved based on the relationship between annotations. Therefore, the annotation process here consists of the following two steps: (1) The first annotation, namely the category annotation, can be achieved using the maximum vote entropy principle based on the results of optimally trained multi-view classifiers; (2) Other annotations on the annotation list can be successively obtained based on the relationship between annotations.

2.3.1 Annotation Process of the Category Annotation

For each unlabeled image, the maximum vote entropy principle is employed to obtain the first annotation, namely the category annotation. Let l_{qv}^k denote that an unlabeled image x_q from the unlabeled image set $\{x_1, x_2, \dots, x_q, \dots, x_Q\}$ is assigned the annotation k over the view v , and p_{qv}^k has the maximum value in the annotation probability set p_{qv} :

$$p_{qv}^k = \arg \max_k \left(p_{qv}^1, p_{qv}^2, \dots, p_{qv}^k, \dots, p_{qv}^K \right) \quad (6)$$

Vote entropy, as a measure of the level of disagreement, is here utilized to metric the purity of annotation results over all views:

$$VE_q = \sum_{k=1}^K \frac{vt(l_{q(\cdot)}^k)}{V} \log \frac{vt(l_{q(\cdot)}^k)}{V} \quad (7)$$

where $vt(l_{q(\cdot)}^k)$ denotes the number of votes that the unlabeled image x_q is assigned to the annotation k in each view v .

The unlabeled image x_q can then be automatically assigned to the annotation which has a maximum number of votes when the following inequality is true:

$$VE_q \geq \mu - \Phi^{-1}(\alpha) \times \delta \quad (8)$$

where Φ is standard normal distribution of annotation result over all unlabeled images as formulated in Eq. (9), μ and $\sigma\psi$ are the expectation and variance of the standard normal distribution respectively, and $\beta\psi$ is the level of disagreement confidence.

$$VE = \sum_{q=1}^Q VE_q \quad (9)$$

2.3.2 Annotation Process of Other Annotations

After achieving the category annotation, other annotations on the annotation list can then be successively obtained with respect to the correlation between these annotations, which is formulated as follows:

$$corr(y_i, y_j) = \frac{num(y_i, y_j)}{\min(num(y_i), num(y_j))} \quad (10)$$

where $num(y_i)$ and $num(y_j)$ are defined as the number of images annotated by annotation y_i and y_j respectively, and $num(y_i, y_j)$ is defined as the number of images annotated by both y_i and y_j .

3 Data Set and Experimental Setup

In this section, we will first present the dataset used in our experiments, then describe the features selected to train multi-view classifiers, and finally discuss the evaluation measures for image annotation.

3.1 Dataset

All the experiments in this work are conducted on a dataset consisting of 50000 images, crawled from the image share website Flickr. Ten most popular tags, including bird, bear, cat, flower, fox, plane, tree, train, sky, and sunset, are selected to implement the annotation-based retrieval. The associated annotations of each retrieved image are ranked according to the option of interesting, and the top 5000 images for each type of retrieval tag are collected together with their associated information, including tags, uploading time, location, etc. Since many of the collected annotations are misspelling or meaningless, it is necessary to perform a pre-filtering for these annotations. More specifically, one annotation can be kept only when the annotation is matched with a term in the Wikipedia. In our case, 17226 unique annotations in total are obtained.

For quantitative evaluation, 24000 images randomly selected from the dataset are adopted as the training set and other 26000 images are selected as the testing set. To get the ground truth, ten volunteers are invited to view each image and exhaustively give their own annotations. Then, the ground truth annotations of each image are the intersection of the given annotations.

3.2 Feature Selection

The problem of feature selection has been an active research topic for many decades due to the fact that feature selection might have a great impact on the final annotation results. In the implementation, the following low-level features including color, texture, semantic description, and textual feature are extracted as the visual descriptors:

- 128-D histogram in HSV color space with 8 bins for Hue and 4 bins each for Saturation and Value;
- 44-D auto-correlogram in HSV color space with inhomogeneous quantization into 44 color bands;
- 225-D block-wise moments in LAB color space, which are extracted over 5×5 fixed grid partitions, each block is described by a 9-D feature;
- 18-D edge direction histogram in HSV color space with two 9-dimensional vectors in horizontal and vertical directions respectively;
- 36-D Pyramid Wavelet texture extracted over 6-level Haar Wavelet transformation, each level is described by a 6-D feature: mean and variance of coefficients in high/high, high/low, and low/high bands;
- A bag-of-words representation obtained using the scale-invariant feature transform method;

- A set of textual feature extracted from annotations associated to each image and a bag-of-words representation obtained using a tf.idf term weighting after filtering out those rare words.

3.3 Evaluation Measures

An annotation-based evaluation measure, normalized discounted cumulative gain at the top s ($NDCG@s$ for short), is here utilized to evaluate the annotation performance. In contrast to other measures, such as precision and recall which only measure the accuracy of annotation results, $NDCG@s$ can measure different levels of relevance and prefer optimal relevance as close as possible to actual relevance. Therefore, this evaluation metric can be utilized to better reflect users' requirement of ranking the most relevant annotations at the top of the annotation list, and is computed as:

$$NDCG@s = \frac{1}{\Gamma} \sum_{i=1}^s \frac{2^{rel(i)} - 1}{\log(1+i)} \quad (11)$$

where Γ , a normalization factor derived from the top s annotations on the annotation list, is here adopted to ensure the value of $NDCG@s$ is 1, and $rel(i)$ is the relevance of an assigned annotation at the location i .

Here, the relevance of each assigned annotation is divided into the following five levels: the most relevant is set as 5, relevant is set as 4, partially relevant is set as 3, weakly relevant is set as 2, and irrelevant is set as 1. The $NDCG@s$ of each unlabeled image is first computed, and then the mean of $NDCG@s$ over the whole unlabeled images are here reported as the final metric of performance evaluation.

3.4 Compared Algorithms

In the current implementation, the proposed multi-view semi-supervised classification approach is compared with other three annotation approaches: one supervised view-specific classification approach (SVS for short) [16], one semi-supervised view-specific classification approach (SSVS for short) [17], and one semi-supervised multi-view classification approach (SSMV for short) [18].

4 Experimental Results

In this section, we will first introduce the process of optimal training over each view-specific classifier, then present the experimental results and give some analysis.

4.1 Optimal Training Process

To simulate a semi-supervised learning scenario, the training dataset is manually divide into two subsets: one subset is utilized as the labeled set in which the labels are known,

and the other subset is utilized as the pseudo-labeled set in which the labels are hidden. The split is performed randomly, and each experiment is repeated 50 times, over 50 different subsamplings of the training data. Four different split size of the labeled set are 1000, 2000, 5000, and 10000 respectively, and the corresponding split size of the pseudo-labeled set are 23000, 22000, 19000, and 14000 respectively. In each split, the proportion of each class is kept similar to what is observed in the training dataset. The test collection is left unchanged, and all reported experimental results apply to this test collection.

4.2 Experimental Results

Performance comparison with respect to average $NDCG@s$ of four different annotation methods with 1000 labeled and 23000 pseudo-labeled samples are illustrated in Fig. 2, where the experimental results are averaged over 50 random splits of the training samples and all the classes. It can be clearly seen from Fig. 2 that the proposed semi-supervised multi-view method outperforms the supervised view-specific annotation method SVS, the semi-supervised view-specific annotation method, as well as the semi-supervised multi-view annotation method SSMV. This improvement confirms the method from the following two aspects: one is that the performance of multi-view classifiers can be improved by adding pairs of pseudo-labeled images with high confidence into the initial labeled image set to iteratively retraining; the other is that the property of multi-view features can be utilized in the process of training stage and testing stage to improve the annotation performance.

The performance comparison between different methods with respect to different labeled samples is shown in Fig. 3, which shows the similar conclusions as in Fig. 2.

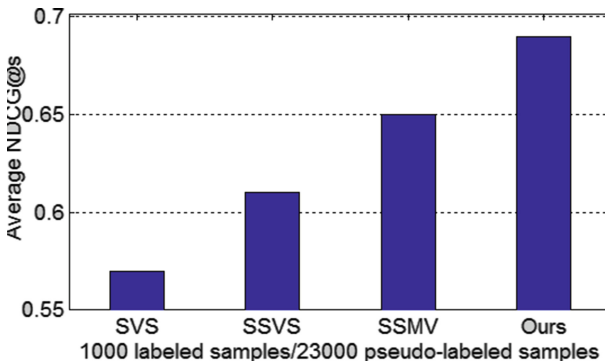


Fig. 2. Performance comparison of different annotation methods in terms of average $NDCG@s$ with 1000 labeled and 23000 pseudo-labeled samples.

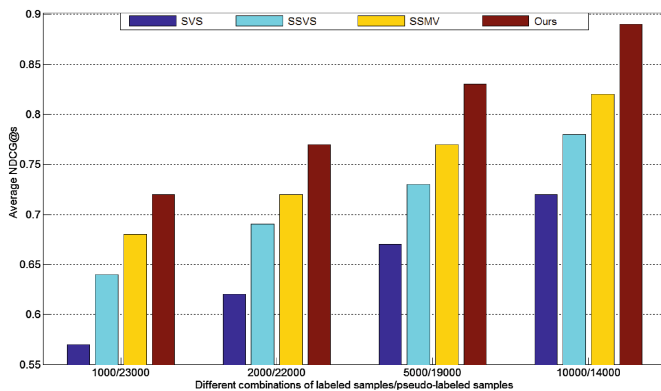


Fig. 3. Performance comparison of different annotation methods in terms of average $NDCG@s$ with different labeled and pseudo-labeled samples.

5 Conclusion

In the training process, each view-specific classifier is first trained independently using uncorrelated and sufficient views, and each view-specific classifier is then iteratively re-trained using initial labeled samples and additional pseudo-labeled samples with high confidence. In the annotation process, each unlabeled image is assigned appropriate semantic annotations based on the maximum vote entropy principle and the relationship between annotations. Experiments conducted on the Flickr dataset show that the proposed scheme can better reflect the images' visual content compared with state-of-the-art methods.

It is worth noting that although the performance of the annotation has been improved to some extent, there are several potential works for future development. Firstly, experiments will be conducted on larger datasets and different datasets using the proposed annotation method. Then, other features will be investigated to improve the performance of image annotation. Finally, we will explore annotation quality improvement problem in a more general scenario, such as annotation categorization, to construct the better lexical indexing for social images.

Acknowledgement. This work is supported by Postdoctoral Foundation of China under No. 2014M550297, Postdoctoral Foundation of Jiangsu Province under No. 1302087B.

References

1. Duygulu, P., Barnard, K., de Freitas, J.F.G., Forsyth, D.: Object recognition as machine translation: learning a lexicon for a fixed image vocabulary. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002, Part IV. LNCS, vol. 2353, pp. 97–112. Springer, Heidelberg (2002)

2. Feng, S., Manmatha, R., Lavrenko, V.: Multiple Bernoulli relevance models for image and video annotation. *IEEE Trans. Circ. Syst. Video Technol.* **13**(1), 26–38 (2003)
3. Wang, B., Li, Z.: Image annotation in a progressive way. In: *Proceedings of the IEEE Conference on Multimedia and Expo*, pp. 811–814 (2007)
4. Tang, J., Li, H., Qi, G.: Integrated graph-based semi-supervised multiple/single instance learning framework for image annotation. In: *Proceedings of the ACM Conference on Multimedia*, pp. 631–634 (2008)
5. Wang, C., Blei, D., Li, F.: Simultaneous image classification and annotation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1903–1910 (2009)
6. Tang, J., Chen, Q., Yan, S.: One person labels one million images. In: *Proceedings of the ACM Conference on Multimedia*, pp. 1019–1022 (2010)
7. Ulges, A., Worring, M., Breuel, T.: Learning visual contexts for image annotation from flickr groups. *IEEE Trans. Multimedia* **13**(2), 330–341 (2011)
8. Verma, Y., Jawahar, C.V.: Image annotation using metric learning in semantic neighbourhoods. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) *ECCV 2012, Part III*. LNCS, vol. 7574, pp. 836–849. Springer, Heidelberg (2012)
9. Zhu, S., Hu, J., Wang, B., Shen, S.: Image annotation using high order statistics in non-euclidean spaces. *Vis. Commun. Image Represent.* **24**(8), 1342–1348 (2013)
10. Huang, J., Liu, H., Shen, J., Yan, S.: Towards efficient sparse coding for scalable image annotation. In: *Proceedings of the ACM Conference on Multimedia*, pp. 947–956 (2013)
11. Marin-Castro, H., Sucar, L., Morales, E.: Automatic image annotation using a semi-supervised ensemble of classifiers. In: Rueda, L., Mery, D., Kittler, J. (eds.) *CIARP 2007*. LNCS, vol. 4756, pp. 487–495. Springer, Heidelberg (2007)
12. Sayar, A., Vural, F.: Image annotation with semi-supervised clustering. In: *Proceedings of the IEEE Symposium on Computer and Information Sciences*, pp. 12–17 (2009)
13. Liu, W., Tao, D., Cheng, J., Tang, Y.: Multiview hessian discriminative sparse coding for image annotation. *Comput. Vis. Image Underst.* **118**(1), 50–60 (2014)
14. Zhang, X., Cheng, J., Xu, C., Lu, H., Ma, S.: Multi-view multi-label active learning for image classification. In: *Proceedings of the IEEE Conference on Multimedia and Expo*, pp. 258–261 (2009)
15. Sridharan, K., Kakade, S.: An information theoretic framework for multiview learning. In: *Proceedings of the Annual Conference on Learning Theory*, pp. 403–414 (2008)
16. Blum, A., Mitchell, T.: Combining labelled and unlabelled data with co-training. In: *Proceedings of the IEEE Conference on Learning Theory*, pp. 92–100 (1998)
17. Zhu, S., Liu, Y.: Semi-supervised learning model based efficient image annotation. *IEEE Sig. Process. Lett.* **16**(11), 989–992 (2009)
18. Amini, M., Usunier, N., Goutte, C.: Learning from multiple partially observed views—an application to multilingual text categorization. In: *Proceedings of the IEEE Conference on Neural Information Processing Systems*, pp. 28–36 (2009)