Discovering Phrase-Level Lexicon for Image Annotation

Lei Yu[1], Jing Liu[1], and Changsheng Xu[1,2]

[1]Institute of Automation, Chinese Academy of Science,
95 Zhongguancun East Road, 100190 Beijing, China
[2]China-Singapore Institute of Digital Media,
21 Heng Mui Keng Terrace, 119613 Singapore
{lyu,jliu,csxu}@nlpr.ia.ac.cn

Abstract. In image annotation, the annotation words are expected to represent image content at both visual level and semantic level. However, a single word sometimes is ambiguous in annotation, for example, "apple" may refer to a fruit or a company. However, when "apple" combines with "phone" or "fruit", it will be more semantically and visually consistent. In this paper, we attempt to find this kind of combination and construct a less ambiguous phrase-level lexicon for annotation. First, concept-based image search is conducted to obtain a semantically consistent image set (SC-IS). Then, a hierarchical clustering algorithm is adopted to visually cluster the images in SC-IS to obtain a semantically and visually specific image set (SVC-IS). Finally, we apply a frequent itemset mining in SVC-IS to construct the phrase-level lexicon and associate the lexicon into a probabilistic annotation framework to estimate annotation words of any untagged images. Our experimental results show that the discovered phrase-level lexicon is able to improve the annotation performance.

Keywords: phrase-level lexicon, image annotation, word correlation

1 Introduction

With the advent of digital imagery, explosive growth of images has led to an increasing need for effectively indexing and searching these images. Image annotation is a promising way to this end.

Image annotation is to find suitable concepts (annotation words) which is able to represent the visual content of an untagged image. A lot of methods have been proposed for image annotation by modeling the correlation between the images and concepts over a tagging dataset[1][2][3]. However, little attention has been paid on what kinds of annotation words are appropriate to annotate images. Generally, semantic clarity and visual representativity are important factors and ideal properties for an annotation lexicon. That is, semantically ambiguous words (e.g., "apple" can be a kind of fruit or a company) or visually diverse words (e.g., "Beijing" images can range from busy street scenes to beauty spots in the city) are unsuitable to be annotations. Recently, some researchers

have made efforts to evaluate the tag clarity or the visual representativity. Lu et al.[4] did the pioneering work to identify the concepts with small semantic gap by defining a confidence score to every concept. The concepts with high scores are more visually and semantically consistent. In [5], Sun et al. considered the bag of words represented images as textual documents and extended the notion of clarity score to search the visually representative tags. In text retrieval, clarity score measures the effectiveness of query keyword. Weinberger et al.[6] conducted an interesting work for tag suggestion, in which a probabilistic framework is proposed to evaluate the tag ambiguity of a tag set. Tag pairs that best disambiguate the set are recommend for annotation. Obviously, the delicately chosen concepts can be utilized to bring the improvement of image annotation. However, only a subset of an original annotation lexicon can be exploited in the annotation process, while several ambiguous but meaningful ones are discarded. Actually, a combination of words (denoted as phrase in the rest of paper) is able to disambiguate a single word, even an ambiguous one, e.g., "apple computer" is more specific than either "apple" or "computer".

Motivated by this view, in this paper, we devote to the discovery of less ambiguous phrase lexicon for better representing and further effectively annotating images than single word lexicon. Concept-based image search by querying with each word in original lexicon is first conducted to obtain a semantically consistent image set (SC-IS). Then, a hierarchical clustering algorithm is adopted to visually clustering the images in SC-IS to obtain a semantically and visually specific image set (SVC-IS). Finally, we apply a frequent itemset mining in SVC-IS to construct the phrase-level lexicon and associate the lexicon into a probabilistic annotation framework to estimate annotation words of any untagged images. Our experimental results show that the discovered phrase-level lexicon is able to improve the annotation performance.

The rest of paper is organized as follows, section 2 reviews related work. Section 3 describes the construction of phrase-level lexicon. Section 4 introduces how to integrate the proposed lexicon into existing annotation methods. Experimental results are reported and discussed in Section 5. We conclude the paper with future work in Section 6.

2 Related Work

Extensive research efforts have been devoted to automatic image annotation in recent years. In [7], the automatic image annotation approaches are classified into three categories: classification-based methods, probabilistic modeling-based methods and search-based methods. Classification-based methods treat image annotation as a classification problem[1][8]. Each concept(annotation word) is considered as a unique class label. After training classifiers for each concept, the final annotation of the image is obtained from the top-ranked concepts. Probabilistic-based methods formulates the correlation between the images and annotation words by maximizing their joint probability. The representative work includes Latent Dirichlet Allocation Model (LDA)[9], Cross-Media Relevance

Model (CMRM)[10], Continuous space Relevance Model (CRM)[2], and Multiple Bernoulli Relevance Model (MBRM)[11]. Recently developed search-based methods[12][13][3] annotate the images by searching, which is model-free and can be easily extended to the large scale datasets.

While the performance of existing image annotation methods is not satisfactory, annotation refinement methods are proposed to reestimate the annotation results. Textual relationships between the annotation results are used to rerank the concepts in [14][15][16]. In [7] [17][18], the concepts are reranked by combining the textual relationships and visual similarities between the concepts in the annotation results.

Since the word correlations are neglected in the annotation step, these refinement methods seem to be a remedy to the original annotation results.

3 Phrase-Level Lexicon construction

In this section, we present a novel approach to construct the phrase-level lexicon. As mentioned above, the phrases sometimes are more semantically and visually consistent than single word concepts. Our approach can generate and select these phrases from the combinations of single word.

The framework of constructing the phrase-level lexicon is shown in Fig. 1. The lexicon construction process has three steps: (1) concept based image retrieval, (2) concept constrained image clustering, (2) phrases generation. We will illustrate these three steps respectively in the following subsections.

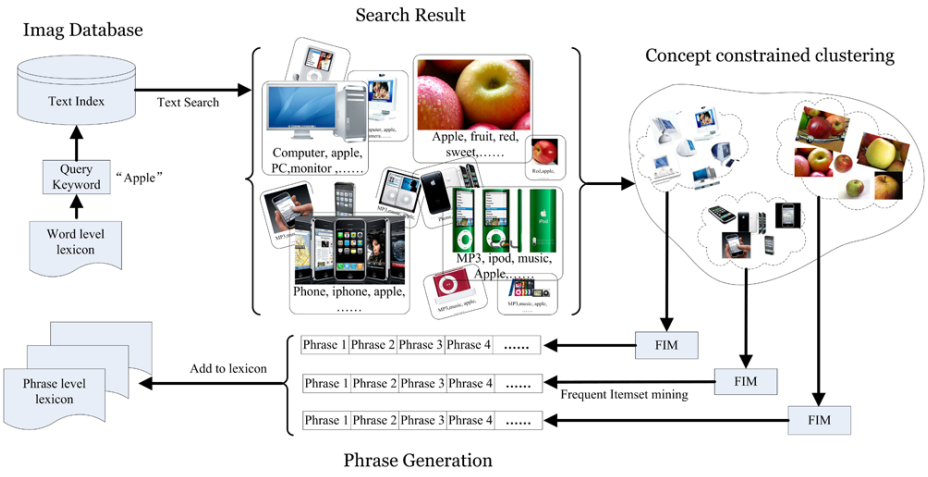

Fig. 1. Framework of Phrase-Level Lexicon Construction

3.1 Concept-Based Image Retrieval

The phrase-level lexicon is constructed based on the original word-level lexicon. Firstly, images and their annotations are collected and stored in the database. Then similar to [3], a keyword based search engine is built which can return the images annotated by the searched keyword. Therefore when one word is put into the search engine, we can obtain a set of images which are semantically consistent to some extent.

3.2 Concept Constrained Image Clustering

Suppose the annotated image set is $\Upsilon = \{T_1, T_2, \ldots, T_i, \ldots, T_{\mathrm{N}}\}$ and the original lexicon space is $\Omega = \{\omega_1, \omega_2, \ldots, \omega_i, \ldots, \omega_{\mathrm{M}}\}$. For every word ω_i in the original lexicon Ω, we get the search result $\mathrm{R}(\omega_i)$. The next step is to cluster the images in $\mathrm{R}(\omega_i)$. Since images in $\mathrm{R}(\omega_i)$ share the same annotation word ω_i, we call this step concept constrained clustering. Then we use the hierarchical agglomerative clustering[19] because it can do clustering only if the similarity matrix of image set is known. The algorithm is described in (Table. 1). After clustering, the visually consistent images are gathered together.

Table 1. Concept Constrained Clustering

Algorithm 1 :	
-Input:	
1. Images in the search result $\mathrm{R}(\omega_i)$;	
2. The final number of clusters K;	
-Output:	
1. The cluster set $\{\mathrm{C}_1, \mathrm{C}_2, \ldots, \mathrm{C}_j, \ldots, \mathrm{C}_K\}$;	
(1)	every image was initialized as a cluster C_j
(2)	while the number of clusters is greater than K
(3)	calculate the distance between two clusters;
(4)	merge the two clusters with the minimal distance;
(5)	end

3.3 Phrase Generation

After the images in the search results $\mathrm{R}(\omega_i)$ are well clustered, for every cluster C_j, we perform a frequent itemset mining(FIM) algorithm[20] on the annotations of images(Table. 2).

The phrase is frequent when its occurence number exceeds the support threshold ε. k-phrase set denotes that every phrase in the set has the length of k word(s). Frequent k-phrase set means that every phrase in the set is frequent and consists of k word(s). Frequent k-phrase set is a subset of k-phrase.

At first, the frequent 1-phrase set is extracted. Then by randomly combining two items in the frequent 1-phrase set into a 2-phrase set, we get frequent

Table 2. Phrases Generation by Frequent Itemset Mining

Algorithm 2 :

-Input:

1. Annotation set $\mathrm{D} = \{d_1, d_2, \ldots, d_i, \ldots, d_s\}$: $s = |\mathrm{C}_j|$ is number of images in cluster C_j, d_i is the annotations of i-th image in C_j;
2. Threshold ε for supporting "frequent";

-Output:

1. The phrases set L;

(1) L_1 = find frequent 1-phrase set(D)
(2) for(k=2, $\mathrm{L}_{k-1} \neq \phi$, k++)
(3) $\quad \mathrm{Z}_k$ = find k-phrase set(L_{k-1})
(4) $\quad$ for each item $z \in \mathrm{Z}_k$
(5) $\quad\quad z.count$ = count occurences(z,D)
(6) $\quad$ end
(7) $\quad \mathrm{L}_k = \{z \in \mathrm{Z}_k | z.count \geq \varepsilon\}$
(8) end
(9) return $\mathrm{L} = \cup_k \mathrm{L}_k$

2-phrases by eliminating the phrases infrequent in 2-phrase set. Hereafter, the combination can be continued to generate larger phrases. Note that when generating $(k+1)$-phrase set from k-phrase($k \geq 2$) set, not all the combinations are valid, only the phrases with length of $(k+1)$ are left.

On one hand, the cluster C_j can be regarded as a visually and semantically consistent subset of $\mathrm{R}(\omega_i)$ and it has less ambiguity than clustering on the whole image set. On the other hand, some noise phrases are eliminated through frequent itemset mining on the cluster. For example, "fruit food" may be a frequent phrase in the set, but they will not be frequent in one cluster because "fruit" and "food" are not visually consistent.

4 Annotation by Phrase-Level Lexicon

When the phrase-level lexicon $\Psi = \{f_1, f_2, \ldots, f_i, \ldots, f_{\mathrm{L}}\}$ has been constructed, the annotation in word-level lexicon can be mapped into phrase-level annotation. Assuming the original annotation matrix A is :

$$\mathrm{A}(j,k) = \begin{cases} 1 & if\ image\ I_k\ is\ annotated\ by\ \omega_j; \\ 0 & else; \end{cases} \tag{1}$$

We construct mapping matrix $\mathcal{F}$ as:

$$\mathcal{F}(i,j) = \begin{cases} 1 & if\ phrase\ f_i\ contains\ \omega_j; \\ 0 & else; \end{cases} \tag{2}$$

where $i \in \{1, 2, \ldots, \mathrm{L}\}$, $j \in \{1, 2, \ldots, \mathrm{M}\}$, and $k \in \{1, 2, \ldots, \mathrm{N}\}$. Then the mapped annotation matrix B of the original image set cab be calculated as

follows:

$$\mathrm{B} = \mathcal{F} \times \mathrm{A} \tag{3}$$

The image set is now annotated with phrases which can be seen as the enhanced "words" of the image annotations. Any kind of annotation methods can be adopted here. In this paper, we apply the similar methods to [11].

Denoting Υ is the training set of annotated images, Ψ is the phrase annotation vocabulary and T_j is an element in Υ. Annotating one image is viewed as a generative process.

We attempt to model the joint probability of observing one annotated image T_j represented by regions $\mathbf{r}_j = \{r_1, \ldots, r_{n_j}\}$ and annotations $\boldsymbol{\psi}_j = \{\psi_1, \ldots, \psi_{m_j}\}$. Since we do not know that $\mathbf{r}_j$ and $\boldsymbol{\psi}_j$ correspond to which image, we calculate the expectation on Υ. Supposing T_j is picked from Υ by probability $P_\Upsilon(T_j)$, the phrase annotations $\boldsymbol{\psi}_j$ are generated by T_j following the independent Multiple-Bernoulli distribution $P_\Psi(\cdot|T_j)$. Instead of modeling the image regions $\mathbf{r}_j$, we assume that T_j produces the real valued feature vectors $\mathcal{G} = \{g_1, \ldots, g_{n_j}\}$ by distribution $P_\Psi(\cdot|T_j)$ and then $\mathbf{r}_j$ is generated by $\mathcal{G}$. $\mathcal{G}$ can be seen as the region-based visual features of the image. Finally, the joint probability of $\{\mathbf{r}_j, \boldsymbol{\psi}_j\}$ is given by:

$$P(\mathbf{r}_j, \boldsymbol{\psi}_j) = \sum_{T_j \in \Upsilon} \left\{ P_\Upsilon(T_j) \prod_{a=1}^{n_j} P_\mathcal{G}(g_a|T_j) \times \prod_{\psi \in \boldsymbol{\psi}_j} P_\Psi(\psi|T_j) \times \prod_{\psi \notin \boldsymbol{\psi}_j} (1 - P_\Psi(\psi|T_j)) \right\} \tag{4}$$

For the untagged image, the region-based visual features $\mathbf{r}_j$ are extracted. Then the annotations $\boldsymbol{\psi}^*$ of the image are the phrase annotations by maximizing:

$$\boldsymbol{\psi}^* = \arg \max_{\boldsymbol{\psi} \in \{0,1\}^\Psi} \frac{P(\mathbf{r}_j, \boldsymbol{\psi})}{P(\mathbf{r}_j)} \tag{5}$$

5 Experiments

In order to evaluate the performance of the phrase-level lexicon, we use the Corel dataset provided by [21] without modification. Corel dataset is widely used in the research of image annotation. The dataset contains 5,000 images from 50 Corel Stock Photo CDs. Each CD has 100 images on the same topic. Each image is segmented into 1∼10 blobs by normalized cut. Every blob is represented by a 36-D visual feature including 18-D color features, 12-D texture features and 6-D shape features. There are 374 annotation words at all and every image is annotated with 1∼5 words.

The dataset is divided into 3 parts: 4,000 training images, 500 evaluation images and 500 images in the test set. The evaluation set is used to find the optimal system parameters. After fixing the parameter, the 4,000 training images and 500 evaluation images are merged into a new training set. This corresponds to the training set of 4500 images and the test set of 500 images used by [21].

To compare the annotation performance with the baseline, a voting process is performed to translate the phrase annotations into words. We partition the

phrase annotations of the image into words and count each word's occurence number. Then the words with top votes are remained as the final annotation. The number of final annotation words of each image is fixed to 5. Because the annotations of images in Corel 5000 dataset are sparse, we only generate the frequent phrases combined by 2 words in the phrases generation procedure. Furthermore, the number of clusters K and the phrase frequent support threshold ε is set as the optimal value on the evaluation set.

We adopt Precision, Recall and F-meature to evaluate the performance of annotation. Their definitions are as follows:

$$Precision(\omega_i) = \frac{1}{n}\sum_{i=1}^{m}\frac{number\ of\ images\ correctly\ annotated\ with\ \omega_i}{number\ of\ images\ annotated\ with\ \omega_i} \quad (6)$$

$$Recall(\omega_i) = \frac{1}{n}\sum_{i=1}^{m}\frac{number\ of\ images\ correctly\ annotated\ with\ \omega_i}{number\ of\ images\ annotated\ with\ \omega_i\ in\ ground\ truth} \quad (7)$$

$$F - meature(\omega_i) = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (8)$$

We further average the values of the precision and the precision, recall and F-meature respectively over all the 260 words in test set. The results of the best 49 words[21] are also reported.

5.1 Experimental Results

Experiment 1: Phrase-Level Lexicon(PLL) VS. Word-Level Lexicon (WLL) The performances of different annotation algorithms are compared in (Table. 3). We denote our method as "PLL-Anno". CRM[2] and MBRM[11] both adopt the probablistic relevance model between images and words to perform annotation. CRM assumes that the annotation words for one image follow a multinomial distribution while MBRM makes the multiple-Bernoulli assumption. Different from [11], we use the same blob visual features in both CRM and MBRM algorithm. CRM-SC and MBRM-SC improves the performances of methods by taking the words co-occurence into consideration[22].

Since CRM and MBRM make an independent assumption of annotation words, their performance is not well. PLL-Anno outperforms CRM-SC and MBRM-SC as well, which shows the phrases generated by our method utilize the word correlation better than these two methods.

The word's occurence varies significantly in Corel dataset. Some words are used hundreds of time while some appear only in several images. For frequently used words, various visual appearance of the word can be catched by visual clustering. This explains why the result of top 49 words gains more improvement than that of 260 words.

Table 3. Performance of different annotation algorithms on PLL and WLL

Algorithm	CRM	MBRM	CRM-SC	MBRM-SC	PLL-Anno
#words with recall > 0	93	115	119	125	121
Results on all 260 words					
Mean Per-word Precision	0.230	0.182	0.190	0.209	**0.214**
Mean Per-word Recall	0.162	0.212	0.232	0.265	**0.271**
Mean Per-word F-meature	0.190	0.195	0.234	0.234	**0.239**
Results on 49 words with top recall values					
Mean Per-word Precision	**0.698**	0.580	0.626	0.636	0.671
Mean Per-word Recall	0.670	0.717	0.721	0.726	**0.766**
Mean Per-word F-meature	0.634	0.641	0.670	0.678	**0.715**

Experiment 2: Comparison to Phrases Generated Without Image Clustering(PLL-NIC) In this experiment, we compare PLL with the phrase-level lexicon generated without image clustering(PLL-NIC). The phrases in PLL-NIC are generated by text search. For every word in the original lexicon, we coupled it with another one word into a phrase. The top k frequently phrases are selected and put into PLL-NIC and the value of k decides the size of lexicon. The annotation results by PLL and PLL-NIC are listed in (Table. 4). We can see that PLL is much more efficient for annotation than PLL-NIC. To gain the same performance, the lexicon size of PLL-NIC is nearly four times than that of PLL. This shows that the visually ambiguous phrases have been eliminated after visual clustering. The phrases in PLL are both semantically and visually consistent. When the lexicon size gets too large, the performances of PLL and PLL-NIC both degrade because more noises are carried into the lexicon by the added low-ranked phrases.

Table 4. Performance Comparison on PLL and PLL-NIC

Lexicon	PLL-NIC				PLL		
#Lexicon size	1185	3980	5506	7044	942	1165	1793
#words with recall > 0	124	125	124	125	116	121	124
Results on all 260 words							
Mean Per-word Precision	0.181	0.201	0.208	**0.219**	0.201	0.208	0.214
Mean Per-word Recall	0.254	0.263	0.261	0.257	0.246	0.263	**0.271**
Mean Per-word F-meature	0.211	0.228	0.232	0.236	0.221	0.232	**0.239**
Results on 49 words with top recall values							
Mean Per-word Precision	0.578	0.647	0.664	0.682	0.647	0.663	**0.671**
Mean Per-word Recall	0.668	0.725	0.731	0.731	0.721	0.751	**0.766**
Mean Per-word F-meature	0.625	0.684	0.696	0.706	0.682	0.704	**0.715**

6 Conclusion

In this paper, we have presented a novel approach to construct a phrase-level lexicon through combining single words in the original word-level lexicon. The generated phrases have more specific meanings and more visual consistency than words. We have proposed a framework to automatically generate and select phrases. Firstly, for every word in the original lexicon, a set of images annotated by this word are retrieved. After the image set is well clustered, a frequent itemset mining(FIM) algorithm is performed on every cluster to select the visually and semantically consistent phrases. At last, all phrases are unified to be the phrase-level lexicon. The experimental results have shown advantages of phrase-level lexicon over word-level lexicon.

In the future, we will testify the effectiveness of our methods on Web images. As we known, there are a lot of noises in the annotations of Web images. Moreover, the annotation words of Web images are not always meaningful as well as the generated phrases. Taking this into consideration, we will investigate some criterion to evaluate the syntax validation of the annotation words and the generated phrases.

Acknowledgments. This work was partially supported by the National Natural Science Foundation of China (9092030360723005 and 60903146), and 973 Program2010CB327905

References

1. Cusano, C., Ciocca, G., Schettini, R.: Image annotation using SVM. In: Proceedings of Internet imaging IV, Vol. SPIE. Volume 5304., Citeseer (2004) 330–338
2. Lavrenko, V., Manmatha, R., Jeon, J.: A model for learning the semantics of pictures, Citeseer (2003)
3. Wang, X., Zhang, L., Li, X., Ma, W.: Annotating images by mining image search results. IEEE Transactions on Pattern Analysis and Machine Intelligence **30**(11) (2008) 1919–1932
4. Lu, Y., Zhang, L., Tian, Q., Ma, W.: What are the high-level concepts with small semantic gaps? In: IEEE Conference on Computer Vision and Pattern Recognition, 2008. CVPR 2008. (2008) 1–8
5. Sun, A., Bhowmick, S.: Image tag clarity: in search of visual-representative tags for social images. In: Proceedings of the first SIGMM workshop on Social media, ACM (2009) 19–26
6. Weinberger, K., Slaney, M., Van Zwol, R.: Resolving tag ambiguity. In: Proceeding of the 16th ACM international conference on Multimedia, ACM (2008) 111–120
7. Wang, C., Jing, F., Zhang, L., Zhang, H.: Content-based image annotation refinement. In: IEEE Conference on Computer Vision and Pattern Recognition, 2007. CVPR'07. (2007) 1–8
8. Li, J., Wang, J.: Automatic linguistic indexing of pictures by a statistical modeling approach. IEEE Transactions on Pattern Analysis and Machine Intelligence **25**(9) (2003) 1075–1088

9. Blei, D., Jordan, M.: Modeling annotated data. In: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, ACM (2003) 127–134
10. Jeon, J., Lavrenko, V., Manmatha, R.: Automatic image annotation and retrieval using cross-media relevance models. In: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, ACM (2003) 126
11. Feng, S., Manmatha, R., Lavrenko, V.: Multiple bernoulli relevance models for image and video annotation. In: Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on. Volume 2.
12. Li, X., Chen, L., Zhang, L., Lin, F., Ma, W.: Image annotation by large-scale content-based image retrieval. In: Proceedings of the 14th annual ACM international conference on Multimedia, ACM (2006) 610
13. Wang, X., Zhang, L., Jing, F., Ma, W.: Annosearch: Image auto-annotation by search. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Volume 2. (2006)
14. Jin, Y., Khan, L., Wang, L., Awad, M.: Image annotations by combining multiple evidence & wordNet. In: Proceedings of the 13th annual ACM international conference on Multimedia, ACM (2005) 706–715
15. Wang, C., Jing, F., Zhang, L., Zhang, H.: Image annotation refinement using random walk with restarts. In: Proceedings of the 14th annual ACM international conference on Multimedia, ACM (2006) 650
16. Wang, Y., Gong, S.: Refining image annotation using contextual relations between words. In: Proceedings of the 6th ACM international conference on Image and video retrieval, ACM (2007) 432
17. Jia, J., Yu, N., Rui, X., Li, M.: Multi-graph similarity reinforcement for image annotation refinement. In: 15th IEEE International Conference on Image Processing, 2008. ICIP 2008. (2008) 993–996
18. Liu, D., Hua, X., Yang, L., Wang, M., Zhang, H.: Tag ranking. In: Proceedings of the 18th international conference on World wide web, ACM (2009) 351–360
19. Xu, D., Chang, S.: Video event recognition using kernel methods with multilevel temporal alignment. IEEE Transactions on Pattern Analysis and Machine Intelligence **30**(11) (2008) 1985–1997
20. Han, J., Kamber, M.: Data mining: concepts and techniques. Morgan Kaufmann (2006)
21. Duygulu, P., Barnard, K., De Freitas, J., Forsyth, D.: Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. Computer VisionECCV 2002 (2002) 349–354
22. Liu, J., Wang, B., Lu, H., Ma, S.: A graph-based image annotation framework. Pattern Recognition Letters **29**(4) (2008) 407–415