Using Multiple Segmentations for Image Auto-Annotation

Jiayu Tang
Intelligence, Agents, Multimedia Group
School of Electronics and Computer Science
University of Southampton, Southampton
SO17 1BJ, United Kingdom
jt04r@ecs.soton.ac.uk

Paul H. Lewis
Intelligence, Agents, Multimedia Group
School of Electronics and Computer Science
University of Southampton, Southampton
SO17 1BJ, United Kingdom
phl@ecs.soton.ac.uk

ABSTRACT

Automatic image annotation techniques that try to identify the objects in images usually need the images to be segmented first, especially when specifically annotating image regions. The purpose of segmentation is to separate different objects in images from each other, so that objects can be processed as integral individuals. Therefore, annotation performance is highly influenced by the effectiveness of segmentation. Unfortunately, automatic segmentation is a difficult problem, and most of the current segmentation techniques do not guarantee good results. A multiple segmentations algorithm is proposed by Russell *et al.* [12] to discover objects and their extent in images. In this paper, we explore the novel use of multiple segmentations in the context of image auto-annotation. It is incorporated into a region based image annotation technique proposed in previous work, namely the training image based feature space approach. Three different levels of segmentations were generated for a 5000 image collection. Experimental results show that image auto-annotation achieves better performance when using all three segmentation levels together than using any single one on its own.

Categories and Subject Descriptors

I.5 [**Pattern Recognition**]: Miscellaneous ; H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

1. INTRODUCTION

Image auto-annotation, which automatically labels images with keywords, has been gaining more and more attentions in recent years. It turns the traditional way of content based image retrieval (CBIR) using low-level image features (colour, shape, texture, etc.) as the query, into an approach that is more favorable to people, namely using descriptive words (semantics). Most of the present auto-annotation models predict captions for the whole images [8, 6, 15], while a few are able to attach words to specific image regions [5, 17, 14]. Annotating images at the whole image level does not indicate which part of the image gives rise to which word, so it is not explicitly object recognition. From this point of view, the second form which generates regional captions is of great interest. However, no matter whether the captions predicted are global or regional, many of the annotation methods choose to segment images first in order to capture local information. Considering the massive work load of manual segmentation, most researchers rely on automatic segmentation techniques [4, 13]. Therefore, the effectiveness of segmentation algorithms have considerable influence on the annotation results. Unfortunately, image segmentation is not a solved problem. It is unrealistic to expect a segmentation algorithm to generate precise partitions. Russell *et al.* [12] try to utilize image segmentation and avoid its shortcomings by using multiple segmentations. In this paper, we examine the use of multiple segmentations for image auto-annotation. It is coupled with a previously proposed region based image annotation approach to see if improvement can be gained compared with the use of single segmentation.

1.1 Related Work

There have been many image auto-annotation techniques in the literature, ranging from statistical inference models [1, 5, 8, 9, 6] to semantic propagation models [11, 7]. For example, Duygulu et al. [5] proposed to use the idea of machine translation for image annotation. They first used a segmentation algorithm to segment images into “object-shaped” regions, followed by the construction of a visual vocabulary, which is represented by ‘blobs’. Then, a machine translation model is utilized to translate between ‘blobs’ comprising an image and words annotating that image.

Yang et al. [17] use Multiple-Instance Learning (MIL) [10] to learn the correspondence between image regions and keywords. “Multiple-instance learning is a variation on supervised learning, where the task is to learn a concept given positive and negative bags of instances”. Labels are attached to bags (globally) instead of instances (locally). In their work, images are considered as bags and objects are instances.

Tang and Lewis [14] propose to realise automatic region based image annotation through a training image based feature space. Mappings of both image regions and textual labels into that space are defined. Similar image segments associated with the same objects are clustered together in this feature space, and should also be close to the labels representing the object. The link between image regions and

CIVR'07, July 9–11, 2007, Amsterdam, The Netherlands

words can be discovered from their separation in the feature space.

All the three techniques described above are able to annotate image regions, and are different from those that only annotate the whole images. For example, Jeon et al. [8] proposed a cross-media relevance model that learns the joint probabilities of a set of regions (blobs) and a set of words, instead of the one-against-one correspondence. We argue that, at some level, models like this benefit from the fact that the data-set contains many globally similar images. As illustrated in [15], a simple global feature descriptor based propagation method achieves even better results on the same data-set. Therefore, region based image annotation techniques are of interest in this work.

On the other hand, since a good segmentation plays an important role in the process of region based image annotation, [12] propose to use multiple segmentations to discover objects and their extent in images. They vary the parameters of a segmentation algorithm in order to generate multiple segmentations for each image. They do not expect any of the segmentations to be totally correct, but "the hope is that some segments in some of the segmentations will be correct". Then, topic discovery models from statistical text analysis are introduced to analyze the segments, in order to find the good ones. Their approach managed to find the correct image segments more successfully than using a single segmentation.

1.2 Overview of Our Approach

Inspired by [12]'s work, we propose to incorporate the idea of multiple segmentations into automatic image annotation. Within a large image data-set, the good segments of the same object will share similar visual features, but the bad ones will have random features of their own. As Russell *et al.* said [12] "all good segments are alike, each bad segment is bad in its own way". We hope that by using multiple segmentations, more good segments can be generated (although from different segmentations), and then captured by auto-annotation models in one way or another.

We chose to embed multiple segmentations into the so-called image based feature space model [14]. There are a few reasons to make this choice of model. Firstly, it is a region based annotation method, as different from those that only annotate the whole images. Secondly, it is easy to implement, and achieves relatively good results. Lastly, transfer from single segmentation to multiple segmentations is more straightforward in this model - what needs to be done is just mapping more segments into the space, without changing the structure or dimensionality.

Firstly, each image is segmented automatically at different segmentation levels into several regions. For each region, a feature descriptor is calculated. We then build a feature space, each dimension of which corresponds to a training image from the database. Finally, we define the mapping of image regions and labels into the space. The correspondence between regions and words is learned based on their relative positions in the feature space. Regional labels that are most likely to be correct are chosen for the entire image.

The details of our algorithm are described in Section 2. Section 3 shows experimental results and some discussions. Finally we draw some conclusions and give some pointers to future work.

2. THE ALGORITHM

2.1 Generating Multiple Segmentations

There are very many different automatic image segmentation algorithms. In this work the Normalized Cuts framework [13] is used, following the choice of [12], because it handles segmentation in a global way which has more chance than some approaches to segment out whole objects. In order to produce multiple segmentations, we varied one parameter of the algorithm, namely the number of segments K. Figure 1 shows some examples of segmented images at different levels of segmentation ($K = 4, 6$ and 8). Evidently, some objects (polar bear, pyramid) get better segmentation at a low level (i.e. a small number of segments), while others (zebra, flower) do so at a high level. However, almost all the object get reasonably good segmentation at one of the levels, although not at the same one.

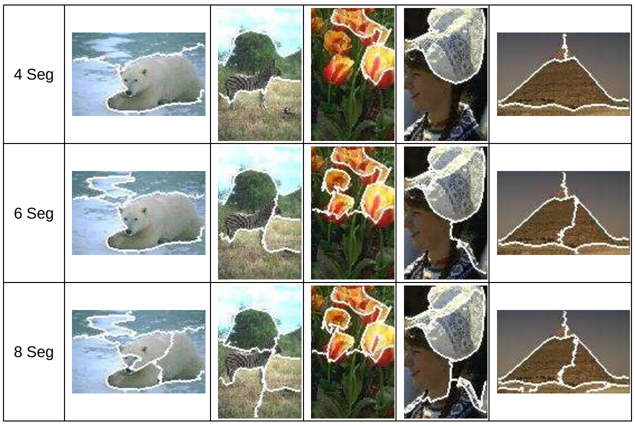

Figure 1: Examples of segmented images at different levels of segmentation

2.2 Incorporating Training Image Based Feature Mapping with Multiple Segmentation

Previous work has shown the effectiveness of a training image based feature mapping in finding representative regions for labels [16], as well as in region based image auto-annotation [14]. However, in both cases a single level of segmentation is used. In this work, we incorporate the image-based feature mapping approach with the idea of multiple segmentations, in order to take into consideration the fact that different objects have their best segmentation at different levels. The details of the modified mapping algorithm is unfolded in the following.

We denote training images as I_i ($i = 1, 2, ...N$, N being the total number of images), and the jth segment in image I_i at the kth segmentation level as I_{ikj}. For the sake of convenience, we line up all the segments from all the segmentation levels of the whole set of training images together and re-index them as I^t ($t = 1, 2, ..., n$, n being the total number of segments). In addition, we denote the vocabulary of the training set as W_l ($l = 1, 2, ..., M$, M being the total number of keywords).

A new mapping $\mathbf{m}$ is defined to map each label and image segment (from all different segmentation levels) into a feature space $\mathbf{F}$. The feature space $\mathbf{F}$ is an N dimensional space where each dimension corresponds to an image from the training set. The coordinate of a label on a particular dimension is decided by the image this dimension represents. If the image is annotated by that label, the coordinate is 1, otherwise it is 0. Specifically, the mapping of word can be defined as

$$\mathbf{m}(W_l) = [e(W_l, I_1), e(W_l, I_2), ..., e(W_l, I_N)] \tag{1}$$

where $e(W_l, I_i)$ indicates whether word W_l exists in training image I_i, which can be further defined as

$$e(W_l, I_i) = \begin{cases} 1 & if\ image\ I_i\ is\ annotated\ with\ W_l \\ 0 & otherwise \end{cases} \tag{2}$$

On the other hand, the coordinates of a segment I^t in $\mathbf{F}$ are defined as:

$$\mathbf{m}(I^t) = [d(I^t, I_1), d(I^t, I_2), ..., d(I^t, I_N)] \tag{3}$$

where $d(I^t, I_i)$ represents the coordinate of segment I^t on the ith dimension, which is either 1 or 0 according to the distance of I^t to image I_i. The distance of a segment to an image is defined as the distance to the closest segment within all the segmentation levels of the image. By comparing segments from all levels, we hope that good segments from different segmentations can be matched, which is less likely when single segmentation is used. The distance between two vectors/histograms V_1 and V_2, which represent the feature descriptors of two segments, is measured by the normalised scalar product (cosine of angle), $cos(V_1, V_2) = \frac{V_1 \bullet V_2}{|V_1||V_2|}$. A threshold t is set to decide if two segments are close enough or not, which then generates either 1 or 0 as the coordinate on one dimension of the space. Mathematically it is defined as follows

$$d(I^t, I_i) = \\ \begin{cases} 1 & if\ max_{k=1,...,m}(max_{j=1,...,n_{ik}}(cos(I^t, I_{ikj}))) > t \\ 0 & otherwise \end{cases} \tag{4}$$

where m is the number of segmentation levels, n_{ik} is the number of segments of image I_i at level k. The mapping of segments can be comprehended as a mapping in which if the object that a segment contains also appears in a particular training image, the coordinate of the segment on the dimension represented by that image is 1, otherwise 0.

We also choose normalised scalar product as the distance measure in space $\mathbf{F}$. Intuitively, segments relating to the same objects or concepts should be close to each other in the feature space. In other words, if in $\mathbf{F}$ the distance of two segments I^x and I^y, which is calculated as $cos(\mathbf{m}(I^x), \mathbf{m}(I^y))$, is very small, they are very likely to contain the same object. Moreover, a label should be close to the image segments associated with the objects the label represents. Suppose the label is W_l, its distance to segments is computed as $cos(\mathbf{m}(W_l), \mathbf{m}(I^t))$.

2.3 Application to Region-Based Image Annotation

To annotate test images, all the test segments from all levels are mapped into the training image based feature space. The test set is denoted as $T_{i'}$ ($i' = 1, 2, ...N'$, N' being the total number of test images), and the j'th segment from the k'th level of image $T_{i'}$ is denoted as $T_{i'k'j'}$. All the test segments are lined up and denoted as $T^{t'}$. By applying the mapping $\mathbf{m}$ to a test segment $T^{t'}$, we can calculate its coordinates in the training image based space as follows

$$\mathbf{m}(T^{t'}) = [d(T^{t'}, I_1), d(T^{t'}, I_2), ..., d(T^{t'}, I_N)] \tag{5}$$

Region based image annotation becomes relatively straightforward once the mapping is done. The probability of a segment being correctly annotated by a particular label, is approximated by their distance in the space. Furthermore, the probability of a test image being correctly annotated by a label, $P(W_l, T_{i'})$, is estimated by the highest probability of this label being correct with any of the segments in that image, as follows

$$P(W_l, T_{i'}) = \\ max_{k'=1,...,m'}(max_{j'=1,...,n_{i'k'}}(cos(\mathbf{m}(W_l), \mathbf{m}(T_{i'k'j'})))) \tag{6}$$

where m' is the number segmentation levels, while $n_{i'k'}$ is the number of segments of test image $T_{i'}$ at level k'. Finally, words with highest value of $P(W_l, T_{i'})$ are chosen as the predicted captions of the image.

2.4 Key Features of the Algorithm

This mapping is similar to the work of [2], in which a region-based feature mapping is used. However, they defined a feature space in which each dimension is an image segment, and then map each image into the space. In other words, the two mappings are essentially the inverse of each other. However, our approach has two main advantages. One is that it is able to map image labels into the feature space, which effectively turns the problem of relating words and regions into one of comparing distances. For [2]'s mapping, there is no way to identify the coordinate of a label on each dimension of the feature space because labels are only attached on an image basis, rather than a region basis. The other is that our approach can be incorporated with multiple segmentations in a more straightforward way. Segments from different levels of segmentations can be mapped into the same space, without making changes to the structure or dimensionality. In contrast, this is not the case for [2]'s approach.

2.5 A Simple Example

In this section a simple example is presented to illustrate the major steps of our algorithm. To make the example more comprehensible and easier, only a single level of segmentation is used here. It is not difficult to transfer the algorithm to the situation of multiple segmentations. The main difference is just that more segments need to be taken into account and more distances need to be calculated.

Consider two annotated training images I_1, I_2 and one unannotated test image T_1; I_1 is labelled as "RED, GREEN" and half of the image is red and the other half is green; I_2 is labelled as "GREEN, BLUE" and half is green and the other half is blue; half of T_1 is red and the other half is blue. Assume the segmentation algorithm manages to separate the two colours in each image and segments them into halves in at least one of the segmentations, we will have four training segments, denoted as I^1, I^2, I^3 and I^4, and two test segments, denoted as T^1 and T^2. Using the RGB values as the feature descriptors, the segments can be represented as

$$\begin{array}{lcl} I^1 & = & (255,0,0); \\ I^2 & = & (0,255,0); \\ I^3 & = & (0,255,0); \\ I^4 & = & (0,0,255); \\ T^1 & = & (255,0,0); \\ T^2 & = & (0,0,255);. \end{array} \quad (7)$$

Then we need to map the test segments into the feature space, which is a two dimensional space in this case as there are two training images. By applying Equation 3, the coordinates of the test segments are as follows:

$$\begin{array}{ll} T^1: & [1,0]; \\ T^2: & [0,1]; \end{array} \quad (8)$$

In addition, the labels can also be mapped into the feature space to give:

$$\begin{array}{ll} RED: & [1,0]; \\ GREEN: & [1,1]; \\ BLUE: & [0,1]; \end{array} \quad (9)$$

It can now be seen that in the feature space, the closest labels for the test segments (regional label predictions) are:

$$\begin{array}{ll} T^1: & RED; \\ T^2: & BLUE; \end{array} \quad (10)$$

3. EXPERIMENT AND RESULTS

Previous work [14] has already demonstrated that the training image based feature space technique outperforms two other state of the art region based image auto-annotation techniques [5, 17]. Other image auto-annotation techniques were not considered because to the best of our knowledge, none of them is able to annotate image regions.

In this work, we compare the effectiveness of using multiple segmentations for image auto-annotation with that of single segmentation. The same image collection[1], which was used in previous work [5, 17, 16], is adopted for the experiment. The dataset contains 5000 images from 50 Corel Stock Photo CDs, and has been divided into a training set of 4500 images and a test of 500 images. Each image had been annotated manually with 1-5 keywords. We used Normalised Cut [13] and varied the parameter of segment number to generate multiple segmentations for each image. In this work, three levels of segmentation are set, 4, 6 and 8. Therefore, the approaches we are comparing are one multiple segmentation approach (denoted as Multi-Seg), which includes three levels 4, 6 and 8, and three single segmentation ones (denoted as 4-Seg, 6-Seg and 8-Seg).

We follow [5]'s representation of regions, which is a 30 dimensional feature vector, including region average colour, size, location, average orientation energy and so on, as detailed in Table 1. Feature vectors are normalised to Z-Scores for distance measure in the image based space. Specifically, suppose the whole set of training feature vectors are $V_1, V_2, ..., V_n$, n being the total number of training image segments, and $V_i = \{V_{i1}, V_{i2}, ..., V_{i30}\}$, we calculate the Z-Score for the jth dimension of the ith vector as follows

$$Z_{ij} = \frac{V_{ij} - mean(V_{1j}, V_{2j}, ..., V_{nj})}{standard\ deviation(V_{1j}, V_{2j}, ..., V_{nj})} \quad (11)$$

Feature of Region	Dimension
Area	1
Position	2
Boundary/Area	1
Convexity	1
Moment of Inertia	1
Average RGB	3
RGB Stdev	3
Average L*a*b	3
L*a*b Stdev	3
Mean Oriented Energy	12

Table 1: Region features

Approaches	Avg. pr.	Avg. re.
4-Seg	0.103	0.129
6-Seg	0.106	0.128
8-Seg	0.100	0.127
Multi-Seg	0.107	0.139

Table 2: Performance comparison of using multiple segmentations for image auto-annotation with single segmentation

Note that for multiple segmentations, mean and standard deviation are calculated over the feature vectors from all segmentation levels, while for single segmentation, they are calculated within each segmentation level. Feature vectors of the test set are also normalised, using the mean and standard deviation of the training vectors.

In order to find the optimal value for threshold t in Equation (4) for each approach, 500 random images are taken out of the training set for evaluation, by training on the remaining 4000 images. Thresholds with the best performances are chosen for the actual auto-annotation experiment. For each test image, the top 5 labels with the highest values of probability are chosen, according to Equation (6).

The *Mean Per-word Precision and Recall*, as used by previous researchers [5, 8, 6, 3, 17], are adopted for evaluation. Per-word precision is defined as the number of images correctly annotated with a given word, divided by the total number of images annotated with this word. Per-word recall is defined as the number of images correctly annotated with a given word, divided by the total number of images having this word in its ground-truth or manual annotations. Per-word precision and recall values are averaged over the set of test words to generate the mean per-word precision and recall. As shown in Table 2, Multi-Seg achieves the best results. In addition, for each approach, we varied the threshold t from 0.99 to 0.80 with step of 0.01 to make further comparisons. *Keyword Number with Recall>0* and *total correct number of words* are evaluated. A keyword has recall>0 if it is predicted correctly once or more, otherwise not. As shown in Figure 2, Multi-Seg managed to predict the most number of keyword with recall>0 ($t = 0.97$) among all the approaches. Moreover, as shown in Figure 3, Multi-Seg managed to predict more correct words than all the single segmentation approaches. In Figure 4, we present some annotation results of the multiple segmentations based approach.

[1] Available at: http://kobus.ca/research/data/eccv_2002/index.html

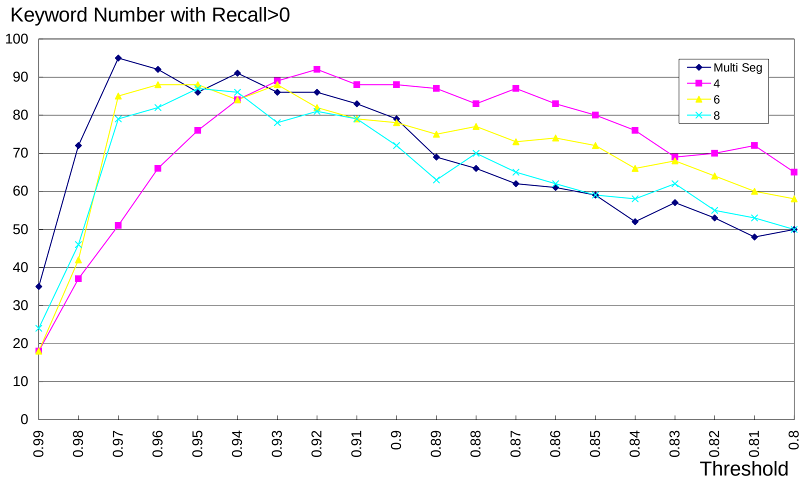

Figure 2: The number of keywords with recall>0 for each approach at different values of threshold

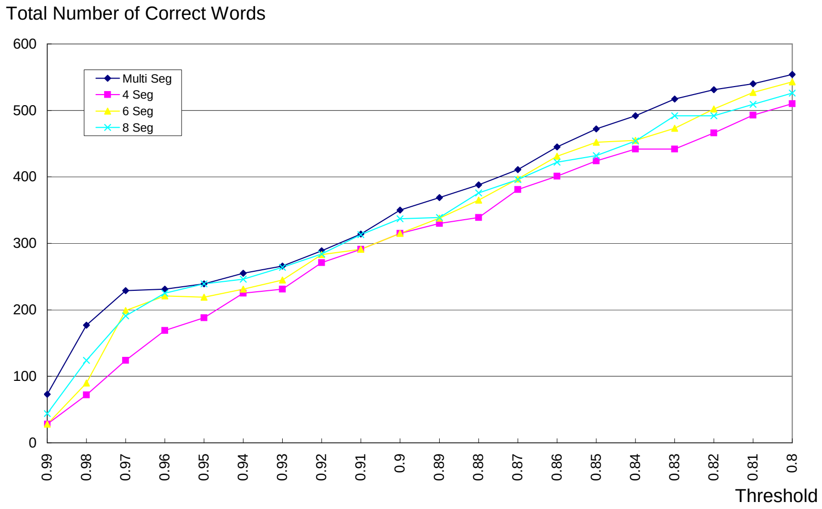

Figure 3: The total number of correctly predicted words for each approach at different values of threshold

4. CONCLUSIONS AND FUTURE WORK

A great number of automatic image annotation techniques use segmentation algorithms to partition the images beforehand. Generally, only one single level of segmentation is chosen, which is assumed to be correct. However, most of the segmentation algorithms do not give satisfying results at this time. We proposed a way of coupling multiple segmentations with image auto-annotation. The parameter of segmentation algorithm is varied to generate several levels of segmentation. On the other hand, a region based image annotation approach, namely the image based feature space, is utilized to incorporate with multiple segmentations. We have shown that annotation performance can be improved on a 5000 image collection when multiple segmentations are used.

As stated in [14], one current disadvantage of the approach is that the feature space has as many dimensions as training images in the set used to build the space. Ways in which the dimensionality of the space can be reduced without losing the association between segments, labels and images is being explored. In addition, the use of a different segmentation algorithm and feature descriptors is planned.

5. REFERENCES

[1] K. Barnard, P. Duygulu, N. de Freitas, D. Forsyth, D. Blei, and M. I. Jordan. Matching words and pictures. *Journal of Machine Learning Research*, 3:1107–1135, 2003.

[2] J. Bi, Y. Chen, and J. Z. Wang. A sparse support vector machine approach to region-based image categorization. In *CVPR '05: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1*, pages 1121–1128, Washington, DC, USA, 2005. IEEE Computer Society.

[3] G. Carneiro and N. Vasconcelos. Formulating semantic image annotation as a supervised learning problem. In *CVPR (2)*, pages 163–168, 2005.

[4] Y. Deng, B. S.Manjunath, and H.Shin. Color image segmentation. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR'99*, volume 2, pages 446–451, Jun 1999.

[5] P. Duygulu, K. Barnard, J. de Freitas, and D. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *The Seventh*

	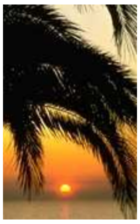	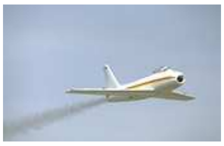	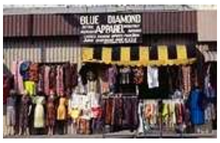	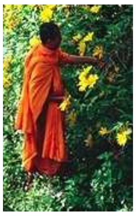
Original	clouds, sun, tree, water	plane, jet, sky	buildings, clothes, shops, street	flowers, garden, monks, people
Multi Seg Annotation	sun, tree, water, jeep, sky	plane, jet, sky, water, snow	people, street, cars, shops, buildings	flowers, people, petals, garden, nest

Figure 4: Some annotation examples by multiple segmentation based approach

European Conference on Computer Vision, pages IV:97–112, Copenhagen, Denmark, 2002.

[6] S. L. Feng, R. Manmatha, and V. Lavrenko. Multiple bernoulli relevance models for image and video annotation. In *Proceedings of the International Conference on Pattern Recognition (CVPR 2004)*, volume 2, pages 1002–1009, 2004.

[7] J. S. Hare and P. H. Lewis. Saliency-based models of image content and their application to auto-annotation by semantic propagation. In *Proceedings of Multimedia and the Semantic Web / European Semantic Web Conference 2005*, 2005.

[8] J. Jeon, V. Lavrenko, and R. Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In *SIGIR '03 Conference*, pages 119–126, 2003.

[9] V. Lavrenko, R. Manmatha, and J. Jeon. A model for learning the semantics of pictures. In *Proceedings of the Seventeenth Annual Conference on Neural Information Processing Systems*, volume 16, pages 553–560, 2003.

[10] O. Maron and T. Lozano-Pérez. A framework for multiple-instance learning. In M. I. Jordan, M. J. Kearns, and S. A. Solla, editors, *Advances in Neural Information Processing Systems*, volume 10. The MIT Press, 1998.

[11] F. Monay and D. Gatica-Perez. On image auto-annotation with latent space models. In *Proceedings of the eleventh ACM international conference on Multimedia*, pages 275–278, 2003.

[12] B. C. Russell, A. A. Efros, J. Sivic, W. T. Freeman, and A. Zisserman. Using multiple segmentations to discover objects and their extent in image collections. In *Proceedings of CVPR*, pages 1605–1614, June 2006.

[13] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, pages 888–905, 2000.

[14] J. Tang and P. H. Lewis. Region based image annotation through a training image based feature space. Submitted to *2007 International Conference on Image Processing (ICIP)*.

[15] J. Tang and P. H. Lewis. Image auto-annotation using 'easy' and 'more challenging' training sets. In *Proceedings of 7th International Workshop on Image Analysis for Multimedia Interactive Services*, pages 121–124, 2006.

[16] J. Tang and P. H. Lewis. An image based feature space and mapping for linking regions and words. In *Proceedings of 2nd International Conference on Computer Vision Theory and Applications (VISAPP)*, Barcelona, Spain, 2007. Accepted.

[17] C. Yang, M. Dong, and F. Fotouhi. Region based image annotation through multiple-instance learning. In *MULTIMEDIA '05: Proceedings of the 13th annual ACM international conference on Multimedia*, pages 435–438, New York, NY, USA, 2005. ACM Press.