# AN IMAGE BASED FEATURE SPACE AND MAPPING FOR LINKING REGIONS AND WORDS

Jiayu Tang and Paul H. Lewis
*Intelligence, Agents, Multimedia Group, School of Electronics and Computer Science*
*University of Southampton, Southampton, SO17 1BJ, United Kingdom*

Keywords: Object Recognition, Image Auto-Annotation.

Abstract: We propose an image based feature space and define a mapping of both image regions and textual labels into that space. We believe the embedding of both image regions and labels into the same space in this way is novel, and makes object recognition more straightforward. Each dimension of the space corresponds to an image from the database. The coordinates of an image segment(region) are calculated based on its distance to the closest segment within each of the images, while the coordinates of a label are generated based on their association with the images. As a result, 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 words can be discovered from their separation in the feature space. The algorithm is applied to an image collection and preliminary results are encouraging.

## 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 image retrieval using low-level image features (colour, shape, etc.) as the query, into an approach that is more favorable to people, namely using words. However, many image auto-annotation techniques only generate labels at the whole image level, rather than at the object or region level. In other words, in this form auto-annotation does not indicate which part of the image gives rise to which word, so it is not explicitly object recognition. Discovering the relationships between image regions and particular textual labels is the problem we wish to tackle in this paper.

### 1.1 Related Work

(Duygulu et al., 2002) view the process of image auto-annotation as machine translation. 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. Thus, it is capable of annotating objects in images.

(Yang et al., 2005) use Multiple-Instance Learning (MIL) (Maron and Lozano-Pérez, 1998) 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." (Maron and Lozano-Pérez, 1998). Labels are attached to bags (globally) instead of instances (locally). In their work, images are considered as bags and objects are instances.

(Russell et al., 2006) 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. Then, topic discovery models from statistical text analysis are introduced to analyze the segments and find 'topics', which correspond to visually similar objects occurring frequently.

Tang J. and H. Lewis P. (2007).
AN IMAGE BASED FEATURE SPACE AND MAPPING FOR LINKING REGIONS AND WORDS.
In *Proceedings of the Second International Conference on Computer Vision Theory and Applications - IU/MTSV*, pages 29-35

## 1.2 Overview of Our Approach

We propose an image based feature space and a mapping of image regions and words into the space for object recognition. Firstly, each image is segmented automatically into several regions. For each region, a feature descriptor is calculated. We then build a feature space, each dimension of which corresponds to an 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.

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 IMAGE SPACE BASED EMBEDDING OF REGIONS AND WORDS

In this section, we first describe how image segments can be represented by visual terms which are based on salient regions. Secondly, we propose how to embed image regions and words into an image based feature space, in order to find the relationships between words and image regions. Then, a simple example is presented as an illustration of the algorithm.

## 2.1 Representing Image Regions by Salient Regions

There are very many different automatic image segmentation algorithms. In this work the Normalized Cuts framework (Shi and Malik, 2000) is used because it handles segmentation in a global way which has more chance than some approaches to segment out whole objects.

Once images are segmented, a descriptor is calculated for each image segment. The approach of (Tang et al., 2006) is followed to represent images by salient regions. Specifically, we first select salient regions by using the method proposed by Lowe (Lowe, 2004), in which scale-space peaks are detected in a multi-scale difference-of-Gaussian pyramid. Lowe's SIFT (Scale Invariant Feature Transform) descriptor is used as the feature descriptor for the salient regions. The SIFT descriptor is a three dimensional histogram of gradient location and orientation. The descriptor is constructed in such a way as to make it relatively invariant to small translations of the sampling regions, as might happen in the presence of imaging noise. Quantisation is applied to the feature vectors to map them from continuous space into discrete space. Specifically, the *k*-means clustering algorithm is adopted to cluster the whole set of SIFT descriptors. Each cluster then represents a visual word from the visual vocabulary. As a result, each image segment can be represented by a k-dimensional frequency vector or histogram, for the visual words contained within the segment.

## 2.2 Image-Based Feature Mapping

We denote images as $I_i$ ($i = 1,2,...N$, N being the total number of images), and the $j$th segment in image $I_i$ as $I_{ij}$. For the sake of convenience, we line up all the segments in the whole set of images together and re-index them as $I^t$ ( $t = 1,2,...,n$, $n$ being the total number of segments).

We define an image-based feature mapping **m**, which maps each image segment into a feature space **F**. The feature space **F** is an $N$ dimensional space where each dimension corresponds to an image from the data-set. The coordinates of a segment in **F** are defined by the mapping **m**:

$$\mathbf{m}(I^t) = [d(I^t,I_1), d(I^t,I_2), ..., d(I^t,I_N)] \quad (1)$$

where $d(I^t,I_i)$ represents the coordinate of segment $I^t$ on the $i$th dimension for which we use 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 the image. Because the number of visual words in a single segment can vary from a few to thousands, the distance between two vectors/histograms $V_1$ and $V_2$, which represent two segments, is measured by normalised scalar product (cosine of angle), $cos(V_1,V_2) = \frac{V_1 \bullet V_2}{|V_1||V_2|}$. Therefore, in this work we define

$$d(I^t,I_i) = max_{j=1,...,n_i} cos(I^t,I_{ij}). \quad (2)$$

Intuitively, segments relating to the same objects or concepts should be close to each other in the feature space.

We can also map labels used to annotate the images into the space. Suppose the vocabulary of the data-set is $W_j$ ($j = 1,2,...,M$, M being the vocabulary size). 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. Therefore, the mapping of words is defined as:

$$\mathbf{m}(W_j) = [d(W_j,I_1), d(W_j,I_2), ..., d(W_j,I_N)] \quad (3)$$

where

$$d(W_j,I_i) = \begin{cases} 1 & if\ I_i\ is\ annotated\ by\ W_j \\ 0 & otherwise \end{cases} \quad (4)$$

Ideally, a label should be close to the image segments associated with the objects the label represents. The normalised scalar product is used to measure the distance between a segment and label, calculated as $cos(\mathbf{m}(I^t), \mathbf{m}(W_j))$.

This mapping is similar to the work of (Bi et al., 2005), 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, one of the advantages of our mapping is that it is also able to map image labels to the feature space. For (Bi et al., 2005)'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. In addition, instead of using global features (colour, shape, texture), we use a histogram of visual words, which are quantised from salient regions within each image segment.

## 2.3 A Simple Example

In this section a simple example is presented to illustrate the major steps of the method. Consider two annotated images; $I_1$ 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. Assume the segmentation algorithm manages to separate the two colours in each image and segments them into halves, we will have four segments in all, denoted as $I^1, I^2, I^3$ and $I^4$. Using the RGB values as the feature descriptors, the segments can be represented as $I^1 = (255,0,0), I^2 = (0,255,0), I^3 = (0,255,0), I^4 = (0,0,255)$. Then we need to map the segments into the feature space, which is a two dimensional space in this case as there are two images. By applying Equation 1, the coordinates of the segments are as follows:

$$\begin{array}{ll} I^1: & [1,0]; \\ I^2: & [1,1]; \\ I^3: & [1,1]; \\ I^4: & [0,1]; \end{array} \tag{5}$$

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

$$\begin{array}{ll} RED: & [1,0]; \\ GREEN: & [1,1]; \\ BLUE: & [0,1]; \end{array} \tag{6}$$

It can now be seen that in the feature space, the closest labels for the segments are:

$$\begin{array}{ll} I^1: & RED; \\ I^2: & GREEN; \\ I^3: & GREEN; \\ I^4: & BLUE; \end{array} \tag{7}$$

# 3 RESULTS AND DISCUSSION

The method has been applied to the Washington image set[1] which contains 697 semantically annotated images. After the original keyword labels were processed by correcting mistakes and merging plurals into singular forms (Hare and Lewis, 2005), the vocabulary consisted of 170 keywords. The whole set of SIFT descriptors are quantized into 3000 visual words. The number of segments is set to 5 per image when using Normalized Cuts (Shi and Malik, 2000). This results in 3241 segments after removing those having no salient regions within them. For each keyword, we find in the feature space the 25 closest segments. The number of correct segments for each keyword is counted manually and those for the 25 keywords (Figure 1) with the highest occurrences in the data-set are reported in Table 1. Because of the fact that the original labels are only attached to the whole image, the decision of whether a segment is correct or not is made by human judgement. We consider a segment being correct if the corresponding object occupies more than 50 percent of the area of the segment, otherwise not. Figure 3 shows some good examples, and Figure 4 shows some bad ones.

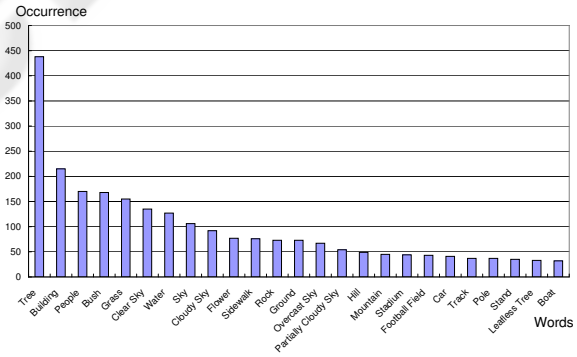



Figure 1: Top 25 Words that appear most frequently in the Washington set.

As shown in Table 1, the results for some keywords are reasonably good, however, for the others they are less so. There are several possible explanations.

[1]Available at: http://www.cs.washington.edu/research/imagedatabase/groundtruth/

1. First of all, some objects are too small in the image to be segmented out reliably by a 5 region N-cut. For example, “Sidewalk”, “Car” and “Boat” usually occupy small areas of the image in the Washington set and rarely achieve good segmentations. Therefore, the algorithm returns the objects that have a high co-occurrence with these words. For “Sidewalk”, image segments with “Trees” are found; For “Car” and “Boat”, segments with “Building” and “Water” are found respectively, as shown in Figure 4.
2. Secondly, some words occur together almost every time they occur and rarely occur separately. This is analogous to an extreme example where a child who has never learnt what a knife and fork look like, is given many images in which both knife and fork appear together, even if he/she is told that all the images contains a knife and fork, there is no way for the child to learn which is which. In the Washington set, “Football Field”, “Track” and “Stand” co-occur almost totally. As shown in Figure 2, for each cell, the number on the dashed line indicates the number of times two words appear together (in the same image), and the other two numbers indicate their occurrence alone without the other. For example, “Track” and “Football Field” occur 36 times together, but only 1 and 7 times respectively on their own. Because of high co-occurrence, the algorithm failed to distinguish them from each other. Almost the same results are returned for them, mostly “Football Field” as shown in Figure 3(c), which is probably because the feature descriptors for “Football Field” are more stable.
3. Lastly, insufficient feature descriptors. Since the SIFT descriptor is using only grey level information, objects that are mainly distinguished by colour will be hard to identify. For example, in this work, the segments returned for “Flower” contain a lot of “Tree” labels (Figure 4(d)), probably because in the data-set, the SIFT feature descriptors for both “Flower” and “Tree” are similar and also often co-occur as well.

Table 1: The number of correct segments out of the top 25 for our method and random choice.

| Keywords | Our Method | Random |
|---|---|---|
| Tree | 21 | 3 |
| Building | 22 | 0 |
| People | 21 | 1 |
| Bush | 25 | 3 |
| Grass | 6 | 1 |
| Clear Sky | 19 | 0 |
| Water | 25 | 2 |
| Sky | 19 | 3 |
| Cloudy Sky | 21 | 3 |
| Flower | 8 | 2 |
| Sidewalk | 2 | 0 |
| Rock | 6 | 2 |
| Ground | 6 | 3 |
| Overcast Sky | 0 | 3 |
| Partially Cloudy Sky | 20 | 4 |
| Hill | 0 | 2 |
| Mountain | 5 | 1 |
| Stadium | 22 | 0 |
| Football Field | 20 | 1 |
| Car | 7 | 1 |
| Track | 2 | 0 |
| Pole | 1 | 0 |
| Stand | 0 | 1 |
| Leafless Tree | 24 | 1 |
| Boat | 0 | 0 |

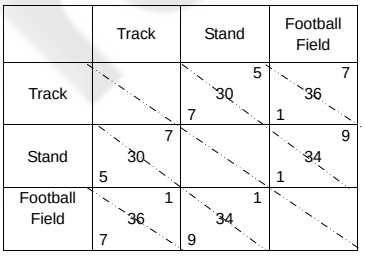


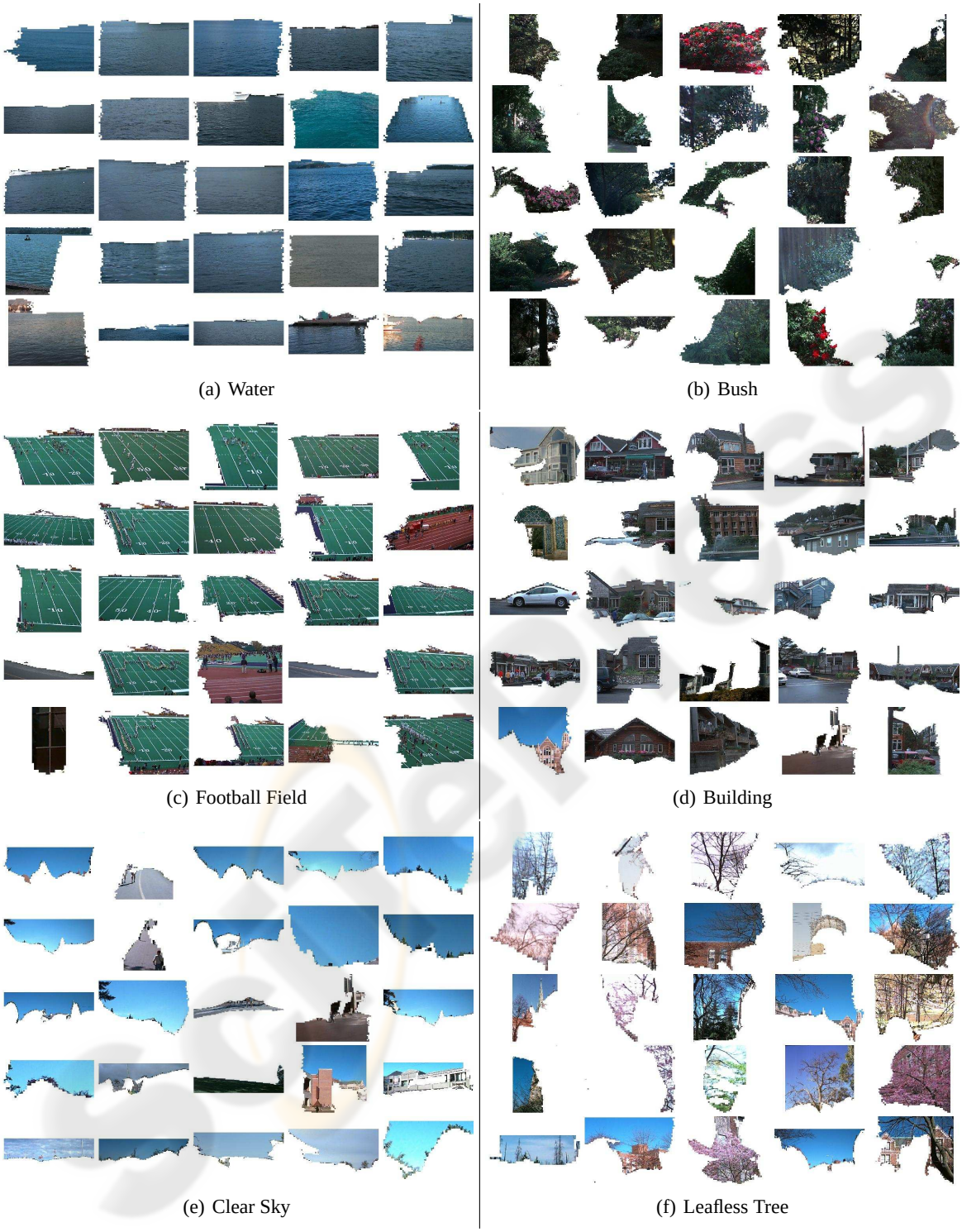

Figure 2: The number of times words “Track”, “Stand” and “Football Field” occur together and separately.

## 4 CONCLUSIONS AND FUTURE WORK

A novel image based feature space has been proposed together with a procedure for mapping in both image segments and textual labels. Some segments associated with the same objects should be clustered together, and also close to the label that represents the object in question. As a result, the relationships between image regions and words can be discovered by comparing their distances in the feature space. Annotating new image segments and images is also straightforward by mapping them into the already built space and finding the closest labels.

One current disadvantage of the approach is that the feature space has as many dimensions as annotated 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 future work, we also plan to investigate the effect of changes in the visual word extraction process and the use of richer feature descriptors than the SIFT. The approach will be tested on other image data-sets

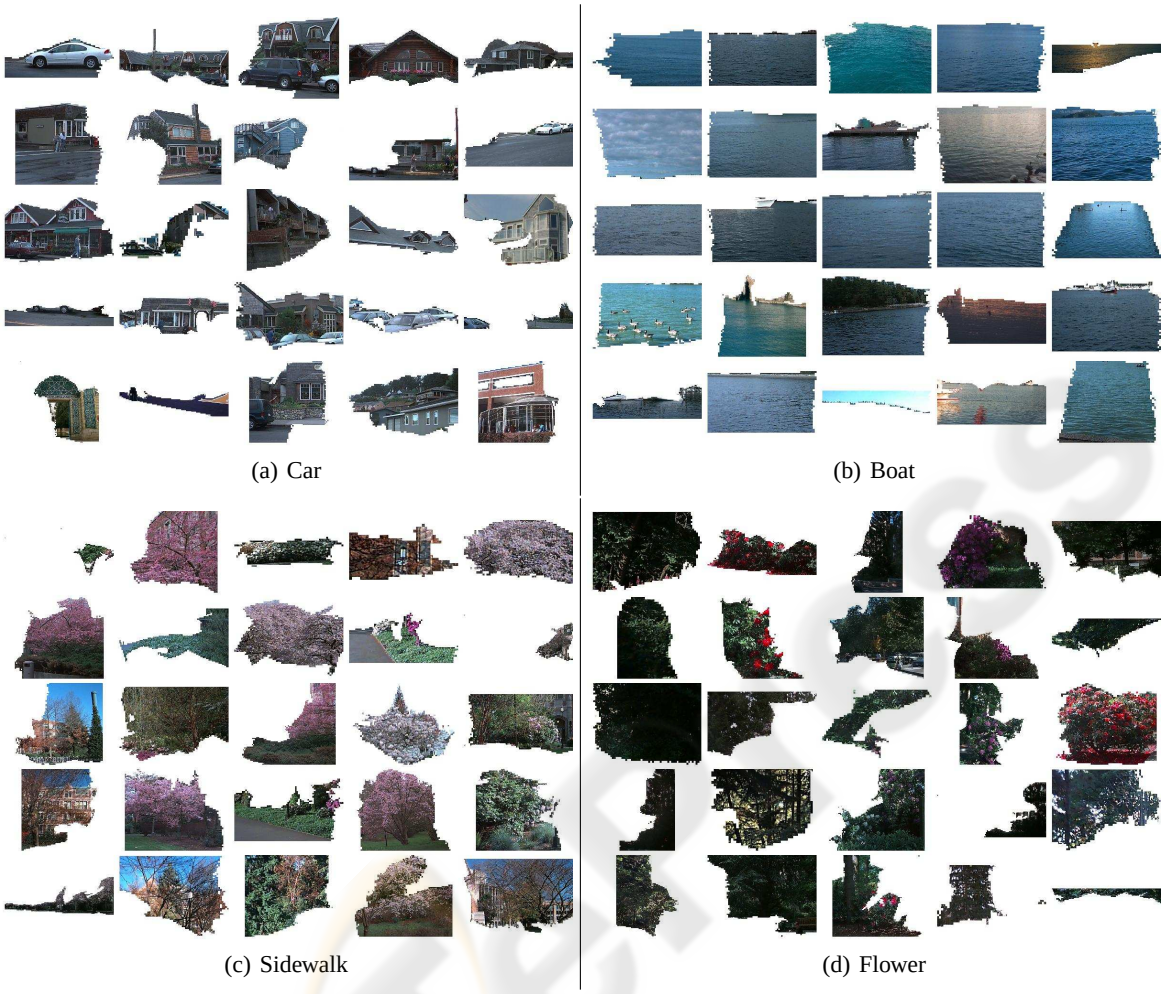


(a) Water (b) Bush

(c) Football Field (d) Building

(e) Clear Sky (f) Leafless Tree

Figure 3: Some good results.



(a) Car (b) Boat

(c) Sidewalk (d) Flower

Figure 4: Some bad results.

for comparison. Since the segmentation is particularly important in this approach, different approaches to segmentation will be evaluated, including in particular, the multiple segmentation idea proposed by (Russell et al., 2006).

## REFERENCES


Bi, J., Chen, Y., and Wang, J. Z. (2005). 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. IEEE Computer Society.

Duygulu, P., Barnard, K., de Freitas, J., and Forsyth., D. (2002). Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *The Seventh European Conference on Computer Vision*, pages IV:97–112, Copenhagen, Denmark.

Hare, J. S. and Lewis, P. H. (2005). 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*.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110.

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

Russell, B. C., Efros, A. A., Sivic, J., Freeman, W. T., and Zisserman, A. (2006). Using multiple segmentations to discover objects and their extent in image collections. In *Proceedings of CVPR*.

Shi, J. and Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*.

Tang, J., Hare, J. S., and Lewis, P. H. (2006). Image auto-annotation using a statistical model with salient regions. In *IEEE International Conference on Multimedia & Expo (ICME)*.

Yang, C., Dong, M., and Fotouhi, F. (2005). 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. ACM Press.