Automatic Image Annotation Using CMRM with Scene Information

Julian Sahertian*[1], Saiful Akbar[2]
[1]Departement of Informatics Engineering, University of Nusantara PGRI Kediri, Kampus 2, Mojoroto Gang I, No.6, Mojoroto, Kediri, Jawa Timur, Indonesia
[1,2]School of Electrical Engineering and Informatics, Bandung Institute of Technology
Jl. Ganesa No. 10, Bandung, Jawa Barat, Indonesia 40132
Corresponding author, email: juliansahertian@unpkediri.ac.id*[1], saiful@informatika.org[2]

Abstract

Searching of digital images in a disorganized image collection is a challenging problem. One step of image searching is automatic image annotation. Automatic image annotation refers to the process of automatically assigning relevant text keywords to any given image, reflecting its content. In the past decade many automatic image annotation methods have been proposed and achieved promising result. However, annotation prediction from the methods is still far from accurate. To tackle this problem, in this paper we propose an automatic annotation method using relevance model and scene information. CMRM is one of automatic image annotation method based on relevance model approach. CMRM method assumes that regions in an image can be described using a small vocabulary of blobs. Blobs are generated from segmentation, feature extraction, and clustering. Given a training set of images with annotations, this method predicts the probability of generating a word given the blobs in an image. To improve annotation prediction accuracy of CMRM, in this paper we utilize scene information incorporate with CMRM. Our proposed method is called scene-CMRM. Global image region can be represented by features which indicate type of scene shown in the image. Thus, annotation prediction of CMRM could be more accurate based on that scene type. Our experiments showed that, the methods provides prediction with better precision than CMRM does, where precision represents the percentage of words that is correctly predicted.

Keywords*: automatic image annotation, CMRM, scene information*

1. Introduction

In the past decade, the digital images have been increasing dramatically with the rapid development of digital cameras and smartphones that can easily capture images. Hence, there is a critical demand for an efficient and effective method that can help users to manage their large volume of image. In order to organize and search images efficiently, content-based image retrieval (CBIR) [1, 2] was proposed. CBIR approach utilizes visual features of the image such as color, texture or shape to obtain relevant images based on user query using query-by-example technique. However, this query technique is less popular for some users. In the other hand, semantic search with a text-based search (eg. document search) still being user's preferred way to search images.

The problem when searching images using text-based search approach is an image that we capture with digital cameras or smartphones doesn't have textual semantic information. To accommodate text-based search, an image requires textual annotation that defines its semantic content information. Image annotation can be done manually or automatically. The manual annotation is not preferable because of involving a lot of efforts. There were several approaches to automatic image annotation e.g. classification approach [3, 4] nearest neighbour approach [5, 6] and statistical approach called relevance model [7-10]. Compared to other methods, relevance model was fairly robust method used in automatic image annotation.

In this paper we propose an automatic annotation method using relevance model and we utilize scene information to improve annotation prediction accuracy. Relevance model assumes that image description pairs are generated by independently chosing visual features for images and words for description. Using a training set of annotated images, joint distribution

Received January 10, 2017; Revised March 28, 2017; Accepted April 13, 2017

of visual features and words is estimated. If we use contextual information of visual features, there is a better chance of predicting correct words for description. One of the contextual information is scene. Human vision can easily identify image scene like '*forest*', '*beach*', etc [11]. Using that scene information, text annotation from detail image objects can be more specifically obtained. For example, object '*tree'* would be more suitable annotation of scene '*forest'*, and object '*sand'* or '*sea'* would be more suitable annotation of scene '*beach'*. Tariq et al. proposed the automatic image annotation model that utilizes scene information [12]. However, they use block-based segmentation for describing image, which is intuitively poor choice for describing objects. Our model is inspired by the model proposed by Tariq et al., but instead of using block-based image regions we use object-shaped regions using automatic image segmentation [13].

This paper is organized as follows. Section 2 explains the image description. Section 3 explains automatic image annotation method. Section 4 explains experiment result and analysis. Section 5 explains conclusion and future works.

2. Image Description

Image is described from local regions using *blobs* representation [7]. To obtain *blobs*, firstly image is divided into several regions. Then, its region visual features such as color, texture, and shape, are extracted and clusterred. Figure 1 shows the process to obtain *blobs*.

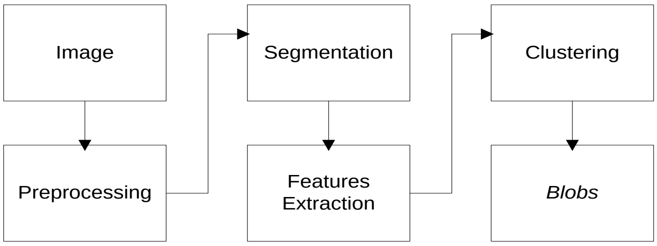


Figure 1. The process of obtaining image *blobs*

2.1. Image Preprocessing

Image preprocessing aims to address lack of contrast between objects and noise problems that cause poor image segmentation result. The problem of lack of contrast between objects is solved using histogram equalization method. Histogram equalization is a method that improves the contrast in an image, in order to stretch out the intensity range. The purpose of histogram equalization is to obtain a wider and more uniform distribution of intensity values histogram. The noise problem is solved by performing median filter to image. Median filter has the same functionality as the mean filter on reducing the noise. However, different from median filter mean filter also reduces the detail of the image which is not a noise.

2.2. Image Segmentation

Segmentation algorithm used in this paper is meanshift color segmentation [13]. Meanshift image segmentation algorithm is similar with k-means image segmentation algorithm. Different from k-means, meanshift does not require the initiation number of clusters and only requires one input parameter. In general, each image has a varying number of different regions, such that for segmenting image meanshift is more suitable than k-means techniques that require the initiated number of segment. Moreover, meanshift is a very robust technique used to image segmentation. L*a*b* color space and Local Binary Pattern [14] image feature is used to get better representation of image pixels. Image segmentation process using meanshift is shown in Figure 2.

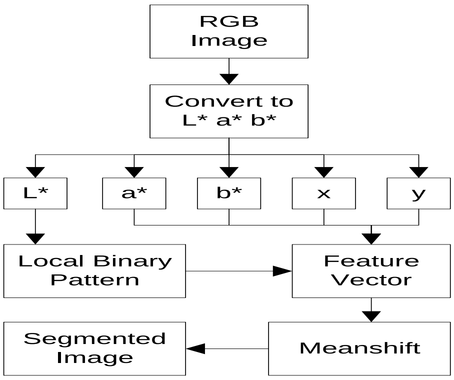


Figure 2. Meanshift image segmentation process

2.3. Feature Extraction

We have implemented a set of features to describe image regions comprising color, texture, and shape. We will describe and discuss the reasons of selecting each image features as the followings.

2.3.1. Color Histogram

Histogram from RGB, Opponent, HSV, and Transformed Color Distribution color space [15] is used to describe color of image regions. The histogram is quantized uniformly in 16 bins and the resulting feature vectorper region has dimension of 48. The histogram is normalized and equalized in order to get invariant property to image scale and illumination.

2.3.2. Local Binary Pattern

Circular uniform Local Binary Pattern [14] is used to describe texture of image regions. There are two parameters for extracting this features that are pixel number and radius. In this reaserach we use 24 pixel numbers and 8 radius based on several experiment conducted. The feature is represented as of histogram of unique label values of each pixel local binary pattern. Similar to the color histogram features, the histogram is normalized in order to get it invariant to image scaling.

2.3.3. Moments

Hu-moments [16] are used to describe the shape of image regions. Hu-moments consist of seven invariant moments. To perform moments feature extraction, firstly we calculate three types of moments: raw moments, central moments, and normalized central moments. Then, we calculate seven invariant moments from normalized central moments.

3. Automatic Image Annotation Method

In this paper, in order to describe objects, we use CMRM method to automatic image annotation [7] instead of using MBRM [9] that uses block-based regions. The following is a brief explanation of the CMRM method we used.

3.1. CMRM Automatic Image Annotation

CMRM is a method of automatic image annotation using relevance model approach. To annotate unlabelled image in this method we estimate joint probability between image region and text annotation. For example, we have the training data of labelled images and test data of unlabelled image. Firstly, each training image is segmented, and then its features such as color, texture and shape are extracted. In order to get blobs representation, the features are clusterred. Thus, each training image in the training dataset is represented as a set of blobs and words. For each unlabelled test images, the image segmentation and blobs generate are also performed. At the training stage, words probability and blobs probability are calculated for each image in the training data T using equations (1) dan (2) [7].

$$P(w|J) = \partial_J \frac{\#(w,J)}{|J|} + \left(1 - \partial_J\right) \frac{\#(w,T)}{|T|} \quad (1)$$

$$P(b|J) = \beta_J \frac{\#(b,J)}{|J|} + \left(1 - \beta_J\right) \frac{\#(b,T)}{|T|} \quad (2)$$

Where, $\#(w,j)$ denotes the number of times word *w* occurs in the caption of *J*, and $\#(w,T)$ denotes the number of times word *w* occurs in the caption of all image in *T*. $\#(b,J)$ is the number of times blob *b* occurs in image *J*, and $\#(b,T)$ is the number of times blob *b* occurs in all image in *T*. $|J|$ is the sum of all the keywords and blobs in image *J*, and $|T|$ is the sum of all the keywords in all image in *T*. ∂ and β are smoothing parameters.

To annotate unlabeled test image is to estimate conditional probability $P(w|I) \approx P(w|b_0, b_1, \dots, b_n)$. Training data *T* is used to estimate the joint probability $P(w, b_0, b_1, \dots, b_n)$ as shown in equation (3) [5].

$$P(w, b_0, b_1, \dots, b_n) = \sum_{J \in T} P(J)\, P(w|J) \prod_{n=1}^{m} P(b_n|J) \quad (3)$$

$P(J)$ is considered to have a uniform distribution for all the images in *T*. The process for obtaining annotation words of unlabeled images using CMRM method is shown in Figure 3.

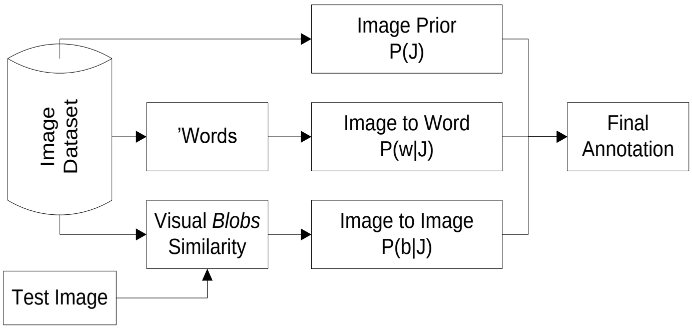


Figure 3. Image annotation using CMRM method

3.2. Scene-CMRM Automatic Image Annotation

In this subsection, we show how to utilize scene information to enhance the prediction accuracy of CMRM. Our automatic image annotation method is called scene-CMRM. Similar to CMRM, in order to annotate unlabeled image using scene-CMRM we estimate the joint probability between image region and text annotation. However, this estimation is calculated based on scene type of that image. The process of image annotation using scene-CMRM is illustrated in Figure 4.

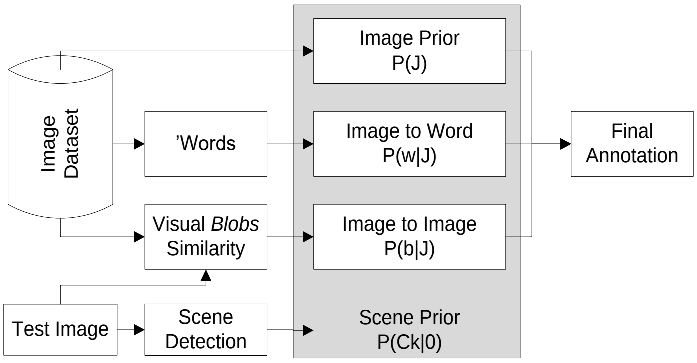


Figure 4. Proposed image annotation method

Mathematical explanation of the proposed automatic image annotation system is as follows. Assume that each image is described from *n* number of visual unit called *blobs* $b = \{b_0, b_1, \ldots, b_n\}$. Each image also has a number of textual descriptions $w = \{w_0, w_1, \ldots, w_n\}$. In addition we assume there are various types of scene $C = \{C_0, C_1, \ldots, C_i\}$ and selection of *blobs* and text annotation depends upon the type of scene at the time of generation of image pairs and their descriptions. Thus, the annotation process of this method is to estimate the joint probability of *b* and *w* conditional on the variable *θ*, where θ is test image scene description. The estimation is shown in equation (4).

$$P(w, b|\theta) = \sum_{C_i \in C} P(C_i|\theta) \sum_{J \in C_i} P(J_i|C_i)\, P(w|J_i) \prod_{n=1}^{m} P(b_n|J_i) \quad (4)$$

$P(w|J)$and$P(b|J)$ from CMRM method changes as shown in (5) dan (6).

$$P(w|J) = \partial_J \frac{\#(w,J)}{|J|} + \left(1 - \partial_J\right) \frac{\#(w,N_i)}{|N_i|} \quad (5)$$

$$P(b|J) = \beta_J \frac{\#(b,J)}{|J|} + \left(1 - \beta_J\right) \frac{\#(b,N_i)}{|N_i|} \quad (6)$$

We replace $|T|$ in the equation (2), which is the sum of all training set is replaced, with $|N_i|$ which is the sum of all image on a cluster scene C_i. The final stage is sorting words in descending order based on the probability value. Text annotations are obtained by taking *n* top number of words with highest probability value or above a certain threshold.

Scene is used as prior knowledge in this proposed automatic image annotation system obtained from scene feature extraction of image. Descriptor is used for scene detection is GIST [11]. GIST feature is able to describe the image scene without going through the process of objects detection in the image. GIST features describe the spatial layout of the image derived from the global feature that is spatial envelope.

After extracting GIST feature, all images are clustered using k-means clustering algorithm to form *k* clusters. For each cluster *i* consist of images *J* represented as J_i with the size of cluster N_i for each scene type C_i. Thus, for each $J \in T$ the probability of selecting image *J* conditional C_i is shown in (7).

$$P(J, C_i) = \begin{cases} \frac{1}{N_i}, & if\ J \in J_i \\ 0, & otherwise \end{cases} \quad (7)$$

For each unlabeled test image, GIST featurea are also extracted from it. This feature is variable θ for the unlabelled image. The probability of selecting scene C_i considered uniform for all scene type.

4. Experiments

This section explains the details of the dataset as well as the experiment scenarios and evaluation measures.

4.1. Dataset

The dataset used for the experiment is image dataset from Microsoft Research in Cambridge (MSRC) [17]. This dataset contains 591 images in 20 categories, and about 80% of the image has more than one label annotation with an average 3 words per image. In total, the vocabulary in this dataset contains 23 words.

In the experiments, image dataset is divided into three parts: 256 images for training, 40 images for evaluation, and 295 images for testing. The evaluation is a process of estimating smoothing parameters. Once the parameters fixed, evaluation data and training data are merged into new training data with 296 images or 50% of the entire dataset. Thus, at the end, the dataset is divided into two parts: 50% for training data and 50% for testing data, as done in [18]. Each image is manually assigned as part of training dataset or testing dataset such that the images are proportionally distributed into the two dataset with respect to their category.

4.2. Evaluation Measures

In order to measure the performance of the proposed method, we used word precision, recall, and f-measure described as the followings.

$$per\ word\ precision(i) = \frac{r_i}{r_i+w_i} \quad (8)$$

$$per\ word\ recall(i) = \frac{r_i}{n_i} \quad (9)$$

$$f - measure = \frac{2\times precision \times recall}{precision+recall} \quad (10)$$

Where r_i is the number of word *i* that correctly predicted, w_i is the number of word *i* that wrongly predicted, and n_i the number of word i in test images.

5. Results and Analysis

In this section we will discuss the experimental results of our automatic image annotation called Scene-CMRM compared to CMRM [7] as the baseline annotation method. For both methods, the experimental results show that the increasing of word prediction increase the precision but decrease recall. The precisions produced by scene-CMRM is slightly better than the precision of CMRM in various number of word prediction as shown in Figure 5. However, the recall of scene-CMRM is less than those of CMRM when the number of word prediction is more than 3 words as shown in Figure 6. This is because the estimation of predicted word in scene-CMRM narrowed down to specific types of scene, where the images and words that are similar are grouped together so the percentage to predict the correct word becomes higher.

As shown in Figure 7, we highlight that the f-measures of both methods achieve the highest value when the number of predicted word per image is 4. As commonly knowns, the f-measure represents the harmonics value between precision and recall. We also note that the f-measure of scene-CMRM is higher than those of CMRM.

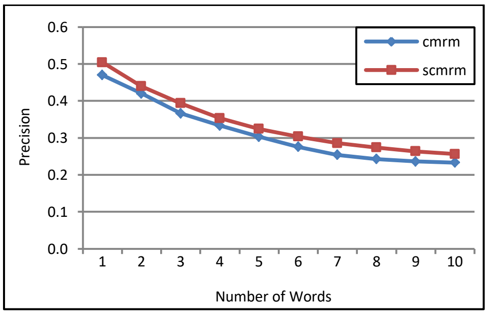


Figure 5. Precision of various number predicted word

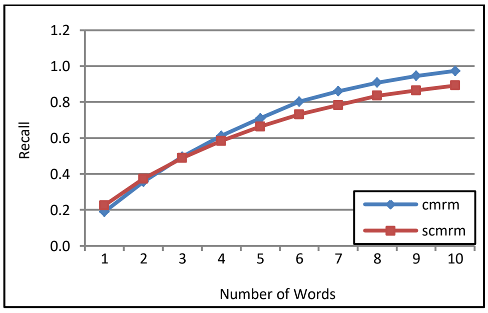


Figure 6. Recall of various number predicted word

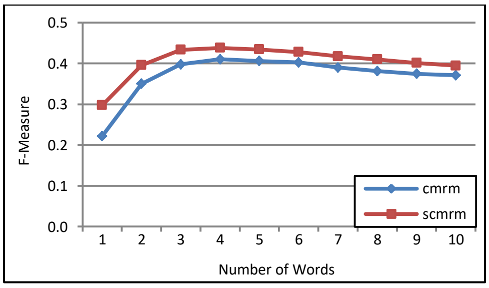


Figure 7. F-Measure of various number predicted word

The detail results of per word precision and recall are shown in Figure 8 and Figure 9.

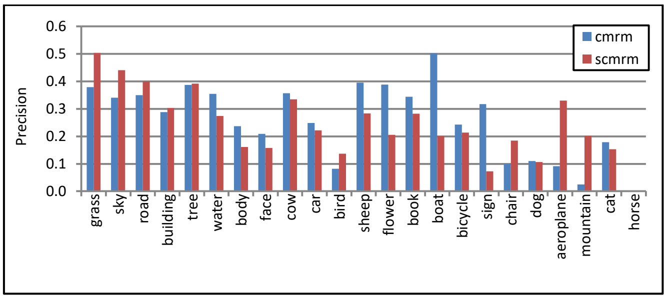


Figure 8. Per word precision

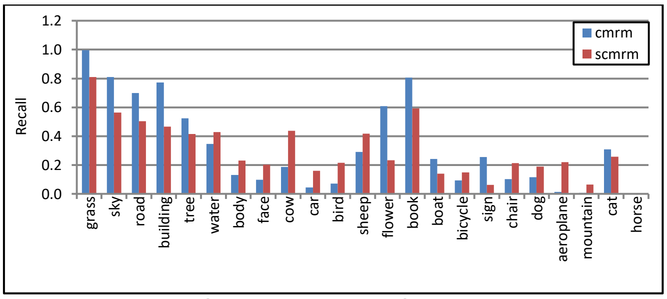


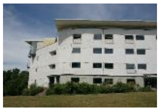
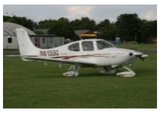
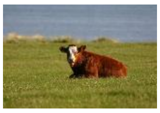
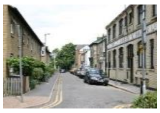
Figure 9. Per word recall

The average of per word precision, recall and f-measure of few words with high frequency on Scene-CMRM and CMRM shown in Table 1. The examples of annotation results from both methods are shown in Table 2.

Table 1. Performance comparison of automatic image annotation method

Method	CMRM [5]	Scene-CMRM
Per word Precision	0.333	0.353
Per word Recall	0.611	0.584
Per word F-Measure	0.410	0.438

Table 2. Annotation examples

Images	Manual Annotation	CMRM [5]	Scene-CMRM
	grass sky building	grass water building sky	grass building water sky
	grass sky building tree aeroplane	grass sky building cat	*tree* building grass sky
	grass water cow	book grass water cow	grass water cow dog
	sky road building tree	building sky road grass	building sky road *tree*

6. Conclusion and Future Works

In this paper we proposed an alternative method to perform automatic image annotation. In the proposed method, we modify the CMRM method by utilizing scene information. Based on the experimental results, the proposed automatic image annotation method called scene-CMRM provides better precision than CMRM but provides worse recall than CMRM. This is because the estimation of predicted word in scene-CMRM narrowed down to specific types of scene, where similar images and words are grouped together so the percentage to predict the correct word becomes higher.

In the future, in addition to the scene information as prior knowledge we can consider the relationship between words by utilizing lexical database like WordNet to enrich words annotation results. Image description can be improved especially in image segmentation to get image regions that close to ideal conditions. The use of more complex visual features like SHIFT should be considered in further works in order to get better representation of regions.

References

[1] Kusrini K, M Dedi I, Ferry Wahyu W. Multi Features Content-Based Image Retrieval Using Clustering and Decision Tree Algorithm. *TELKOMNIKA Telecommunication Computing Electronics and Control*. 2016; 14(4): 1480-1492.

[2] Agus ZA, Rizka WS, Dimas Fanny HP, Dini AN. Region Based Image Retrieval Using Ratio of Proportional Overlapping Object. *TELKOMNIKA Telecommunication Computing Electronics and Control*. 2016; 14(4): 1608-1616.

[3] E Chang, G Kingshy, G Sychay, G Wu. CBSA: Content-Based Soft Annotation for Multimodal Image Retrieval Using Bayes Point Machines. *IEEE Trans. on CSVT*. 2003; 13(1): 26–38.

[4] C Cusano, G Ciocca, R Schettini. *Image annotation using SVM*. Proceedings of the SPIE. San Jose. 2003; 5304: 330-338

[5] Ameesh Makadia, Vladimir Pavlovic, Sanjiv Kumar. Baselines for Image Annotation. *International Journal of Computer Vision*. 2010; 90(1): 88-105.

[6] Xirong Li, CGM Snoek, Marcel Worring. Learning Social Tag Relevance by Neighbor Voting. *IEEE Transactions on Multimedia*. 2009; 11(7): 1310-1322.

[7] J Jeon, V Lavrenko, R Manmatha. *Automatic Image Annotation and Retrieval Using Cross-Media Relevance Models*. Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Toronto, Canada. 2003; 119-126.

[8] Victor Lavrenko, R Manmatha, Jiwoon Jeon. A Model for Learning The Semantics of Pictures. *Advances in Neural Information Processing Systems*. 2003; 16: 553-560.

[9] SL Feng, R Manmatha, V Lavrenko. *Multiple Bernoulli Relevance Models for Image and Video Annotation*. Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Washington DC, USA. 2004; 2: II-1002-II-1009.

[10] Hengame Deljooi, Ahmad R Eskandari. A Novel Semantic Statistical Model for Automatic Image Annotation Using the Relationship between the Regions Based on Multi-Criteria Decision Making. *IJECE International Journal of Electrical and Computer Engineering*. 2014; 4(1): 37-51.

[11] Aude Oliva, Antonio Torralba. Modeling The Shape of The Scene: A Holistic Representation of The Spatial Envelope. *International Journal of Computer Vision*. 2001; 42(3): 145-175.

[12] Amara Tariq, Hassan Foroosh. *Scene-Based Automatic Image Annotation*. IEEE International Conference on Image Processing (ICIP). Paris. 2014; 3047-3051.

[13] Dorin Comaniciu, Peter Meer. Mean Shift: A Robust Approach Toward Feature Space Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2002; 24(5): 603-619.

[14] Timo Ojala, Matti Pietikainen, Topi Maenpaa. Multiresolution Gray-scale and Rotation Invariant Texture Classification with Local Binary Patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2002; 24(7): 971-987.

[15] Koen EA Van De Sande, Theo Gevers, Cees GM Snoek. Evaluating Color Descriptors for Object and Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2010; 32(9): 1582-1596.

[16] Ming-Kuei Hu. Visual Pattern Recognition by Moment Invariants. *IRE Transactions on Information Theory*. 1962; 8: 179-187.
[17] A Criminisi. Microsoft Research Cambridge (MSRC) Object Recognition Pixel-wise Labeled Image Database (Version 2). 2004.
[18] Jiayu Tang. Automatic Image Annotation and Object Detection. Doctoral Thesis. Southampton: University of Southampton, Electronics & Computer Science; 2008.