

Automatic Semantic Topic Discovery Approach of the Line Image Based on Support Vector Machine

Cong Jin, Jin'an Liu

School of Computer Science, Central China Normal University, Wuhan
Email: jinc26@aliyun.com

Received: May 21st, 2014; revised: May 26th, 2014; accepted: Jun. 9th, 2014

Copyright © 2014 by authors and Hans Publishers Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY).
<http://creativecommons.org/licenses/by/4.0/>



Open Access

Abstract

A semantic topic discovery approach of the line image, based on support vector machine, has been proposed in this paper. Firstly, the training images are divided into non-overlapping sub-blocks with same size. After clustering image sub-blocks, we obtained class set generated by cluster centers, and extracted all nouns from text annotation of each training image in order to obtain a keyword set. Secondly, the un-label testing image is also divided into non-overlapping sub-blocks as same as training images, we calculated the correlation between the sub-block and each keyword, and a keywords set for each sub-block may be obtained. Finally, the number of each keyword appearing in the each sub-block is calculated, we let the keywords with maximum to occurrences number be the semantic topics of the line image. The experimental results confirm that proposed automatic semantic topic discovery approach for line image is effective and has good performance.

Keywords

Digital Image, Semantic Topic Discovery, Text Clustering, Support Vector Machine

基于支持向量机的线条图像语义主题自动发现方法

金 聪, 刘金安

华中师范大学计算机学院, 武汉
Email: jinc26@aliyun.com

收稿日期: 2014年5月21日; 修回日期: 2014年5月26日; 录用日期: 2014年6月9日

摘要

提出了一种基于支持向量机分类器的线条图像语义主题自动发现方法。首先对训练图像进行分块, 在对图像子块进行聚类后, 得到由聚类中心构成的类集合; 从每幅训练图像的注释文字中提取所有名词构成关键词集合。其次, 对未标注的测试图像进行同样分块处理, 计算子块与每个关键词的相关性, 得到每个子块的标注词集合。最后, 计算每个关键词在各个子块标注中出现的次数, 取出现次数最多的关键词作为图像的语义主题。实验结果表明, 所提出的方法对于线条图像的语义主题自动发现是有效的, 具有比较好的性能。

关键词

数字图像, 语义主题发现, 图像块聚类, 支持向量机

1. 引言

随着图像数量的急剧增长, 图像已充斥整个网络。如何有效地检索和利用如此大规模的海量图像, 以使用户能够快速、有效地找到感兴趣的图像资源, 已成为当前极具挑战性的任务。然而, 作为图像检索的重要基石, 图像语义标注[1]-[3]的效果决定着图像检索的性能。经过十几年的发展, 图像检索已形成三种主流解决方案: 基于文本的图像检索、基于内容的图像检索和基于语义的图像检索。前两者分别依赖于人工对图像的标注和图像视觉特征的提取。由于人工标注费时费力、个人理解差异大, 使得图像检索出错率高。由于图像视觉特征不能反映图像的内容信息, 导致图像检索性能不理想。基于语义的图像检索[4]可弥补前两种方案的缺陷, 使检索结果与用户所需信息尽可能一致。

根据大量观察我们发现, 人们利用语义检索图像时, 通常通过图像主题来检索图像, 大多数情况下并不利用图像的非主题信息。这启发我们: 如果能发现图像的主题、并通过主题来检索图像, 不仅可以提高图像检索的性能, 而且能迅速聚焦图像的本质信息。基于此, 本文将图像主题的自动发现作为研究内容。

中国古典文学作品中存在大量的插图。这些图大多以线条的形式呈现。线条图像没有颜色特征, 且受当时印刷技术的限制, 图像比较简洁。如果我们能够自动发现这种线条图像的语义主题, 并对所获语义主题进行文本分析与处理, 则可为图像叙事及图像内容理解的实现提供技术支持, 也可为处理和分析海量图像、使用户能够快速有效地定位图像信息奠定基础。

2. 相关的工作

在之前的研究中, 已提出了多种主题发现解决方案[5]-[8]。主题发现也称为主题抽取或主题识别。文献[5]将主题发现分为广义和狭义两种类型。

能够发现代表性信息的方法属于广义主题发现, 其余方法属于狭义主题发现。例如, 在针对文本的主题发现中, 从高频被引论文中抽取高频词来代表主题、基于语义局部性思想来界定主题、以词频变化率处于突发状态的主题词作为主题等多种方式均属于广义主题发现。

从大规模文本中发现主题属于狭义主题发现，其目标是在已经确定了文档集中若干概念的基础上进行主题聚类，以便发现文档中的核心主题。狭义主题发现是目前自然语言处理中的主流主题发现方法。

文献[6]提出的方法能发现文本中的隐含知识，并将其表示为含有主题/副主题的层次树，每个主题包含与其相关的文档集和文档摘要，以便用户从层次树中浏览和选择所需主题。实验结果表明，无论是作为主题检测还是分类和概括工具，该方法都具有较高的效率。

文献[7]提出的基于共词分析的文本主题词聚类与主题发现方法，通过停用词过滤和 TF-IDF 关键词提取技术提取出主题词串，然后通过 K-means 算法对主题词串进行聚类分析，从而发现文本主题。

文献[8]对面向网络论坛的高质量主题发现进行了较为深入的研究。在所提出的通用高质量主题发现框架下，利用特征抽取技术提取内容特征，利用结构特征来发现主题。实验结果表明，所提出的框架能够很好地发现主题，能够在很多 Web2.0 领域得到应用。例如，博客、社会网络平台等。

文献[9]提出一种将潜在主题模型与显著性检测算法融合的图像主题发现方法。该方法首先利用显著性检测算法从图像中识别显著目标，然后通过组合潜在目标与背景部分的子主题，设计与实现一种可以发现图像主题的分层潜在主题模型。

文献[5]-[8]的主题发现方法都是面向文本分析领域，与图像的主题发现没有太大的关系。文献[9]提出了一种图像主题发现方法。虽然文献[9]称为图像主题发现方法，但它与本文的方法不同。首先文献[9]中所说的图像主题指的是图像目标区域，而本文所说的图像主题指的是中文单词，这是本文与文献[9]的根本区别。其次是研究对象不同。本文是针对中国古典文学作品中的线条图像开展语义主题发现研究，这与文献[9]针对彩色图像的图像目标区域研究不同。

3. 支持向量机

分类器模型有很多种，本文采用支持向量机(SVM)实现分类。SVM 是基于统计学习理论[10]的学习方法，通过构造最优超平面，使得对未知样本的分类误差最小。SVM 可以在样本数量很少的情况下达到很好的分类效果。SVM 的基本原理是寻找一个最优分类面，并使其两侧的分类间隔最大。对于非线性可分问题，SVM 首先通过非线性变换，将输入向量映射到高维空间，以使其在高维空间中能够进行线性分类。SVM 对于两类线性可分情形，可直接构造最优超平面，使得样本集中的所有向量满足如下条件：

- (1) 能被某一超平面正确划分；
- (2) 距离该超平面最近的异类向量与超平面之间分类间隔最大。

其中，条件(1)保证经验风险最小，条件(2)使期望风险最小。这里最优超平面的构造问题实质上属于约束条件下的优化问题。一个二次规划问题的最优分类函数为 $f(x) = \text{sgn} \left\{ \sum_{x_i \in SV} \alpha_i^* y_i K(x_i \cdot x) + b^* \right\}$ ，式中 $K(\cdot)$ 是 SVM 的核函数， b^* 是分类阈值。

在分类函数中，某些 x_i 对应的 α_i^* 为零，具有非零值的 α_i^* 对应于支持向量。对于线性 SVM，核函数 $K(\cdot)$ 是两个向量的点积运算；非线性 SVM 则通过非线性映射将输入向量映射到一个高维特征空间来构造最优分类面。选择不同核函数 $K(\cdot)$ ，可以得到不同的支持向量。总之，SVM 首先用内积函数定义的非线性变换将输入空间变换到一个高维空间，在这个空间中求最优分类面。

4. 图像语义主题发现的预处理方法

一般的图像语义标注方法大多建立在图像分割算法的基础上。图像分割算法的效果受分割技术本身以及图像的影响很大，因此本文不采用图像分割，取而代之的是图像分块。

为了自动发现图像语义主题，需要先对图像训练集进行预处理。具体步骤如下。

(1) 对于给定的有标注的图像训练集中的每一幅图像进行大小相等的分块, 获得图像分块集。将所有训练图像的分块集集中在一起, 构成图像训练集的分块集合。

(2) 对所获得的分块集合中的子块进行聚类, 得到由聚类中心构成的类别集合, 记为 $C1 = \{C_1, C_2, \dots, C_n\}$, 其中 n 是类的个数。

(3) 每一幅训练图像都有注释文字说明该图像的语义。本文从每幅图像的注释文字中提取所有的名词构成关键词集合, 记为 $K = \{K_1, K_2, \dots, K_m\}$, 其中 m 是关键词的个数。

(4) 这样, 对于给定的图像训练集, 存在语义标注模板 $M = \{K_1, K_2, \dots, K_m; C_1, C_2, \dots, C_n\}$ 与之对应。

5. 图像语义主题发现过程

在对图像训练集进行预处理后, 利用所获得的语义标注模板可进行图像语义主题发现工作。具体过程如下。

(1) 将未标注的测试图像进行分块, 得到分块子集 $C2 = \{C'_1, C'_2, \dots, C'_n\}$, 其中 n' 是分割子块的个数。

(2) 将集合 $C2 = \{C'_1, C'_2, \dots, C'_n\}$ 的每个子块分类到 $C1 = \{C_1, C_2, \dots, C_n\}$ 的某一类中。

(3) 计算每个关键词 K_i 与 C'_j 的相关性。若 K_i 与 C'_j 的相关性最大, 则将关键词 K_i 标注到 C'_j 上。

(4) 若与 C'_j 有最大相关性的有多个关键词, 则这些关键词均标注到 C'_j 上。

(5) 如此下去, 直到集合 $C2 = \{C'_1, C'_2, \dots, C'_n\}$ 中每个子块均得到自己的标注集。

(6) 统计每个关键词在各个子块标注集中出现的次数, 将出现次数最多的关键词作为图像的语义主题。至此, 可获得图像语义主题自动发现流程图, 见图 1。

6. 图像语义主题发现过程中的实现问题

(1) 子块 C'_j 分类到 C_1, C_2, \dots, C_n 的某一类中

在语义主题发现过程的第 2 步, 需要将每个子块 C'_j 分类到 C_1, C_2, \dots, C_n 的某一类中。在本文中, 利用 SVM 实现分类。

(2) 关键词与分割子块的相关性计算

在语义主题发现过程的第 3 步, 需要计算每个关键词 K_i 与 C'_j 的相关性。本文定义关键词 K_i 与 C'_j 的相关性为 $C(K_i, C'_j)$:

$$C(K_i, C'_j) = \frac{1-\lambda}{m+n'}(\alpha_1 + \alpha_2) + \frac{\lambda}{m+n}(\beta_1 + \beta_2) \quad (1)$$

这里, α_1 是 K_i 在 C'_j 中出现的次数, α_2 是 K_i 在 C'_j 中的系数; β_1 是 K_i 在 C_j 中出现的次数, β_2 是 K_i 在 $C1$ 中出现次数的均值; λ 是调节参数。而 α_2 的计算按下式进行

$$\alpha_2 = \begin{cases} 1, & K_i \text{ 是描述 } C'_j \text{ 中背景的名词} \\ 2, & K_i \text{ 是描述 } C'_j \text{ 中前景的名词} \end{cases} \quad (2)$$

7. 实验与结果

实验用的线条图像均来自中国古代小说中的插图。

在本文中, 选取切磋武艺、芭蕉、伏案写作、树、城墙、倒立、水草、军旗、骑马迎敌、岩石、家俱、囚犯、战马、开会、水井等具有简单主题和复杂主题的各类图像作为实验图像, 每类图像各取 30 幅。取 15 幅图像作为训练样本, 其余 15 幅图像作为测试样本。图 2 是测试图像的实例。

原始图像的大小为 512×640 , 图像子块的尺寸为 64×64 , 参数 $n = 10$, $\lambda = 0.5$ 。

使用 SVM 进行分类时选取哪种核函数是很关键的, 因此本文选取多种核函数在实验样本集中进行仿

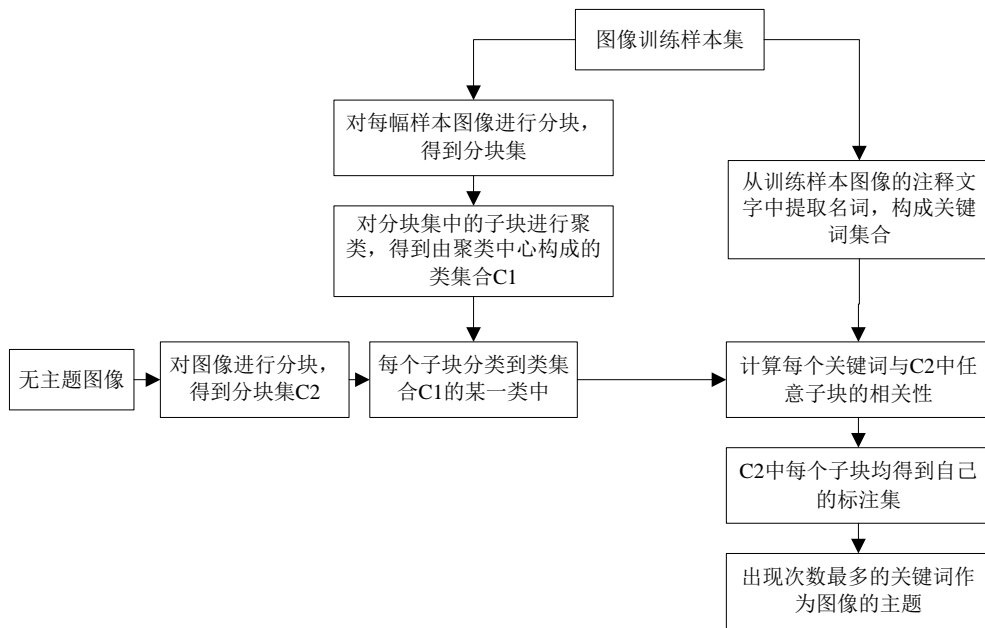


Figure 1. An automatic discovery diagram for image semantic topic

图 1. 图像语义主题自动发现流程图

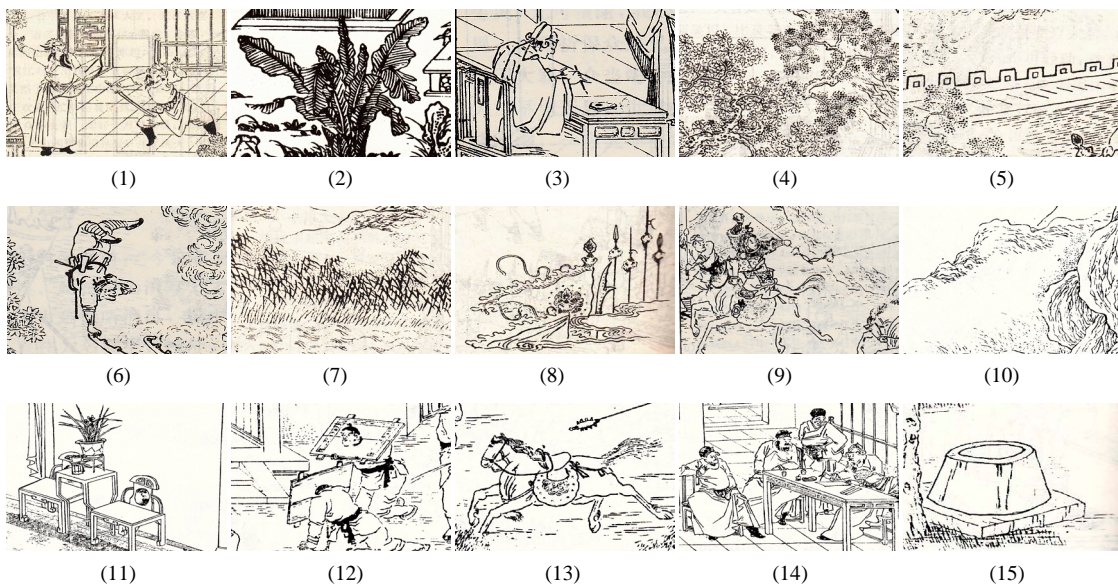


Figure 2. Some image samples

图 2. 实验用测试图像的实例

真实验。当 SVM 的核函数为径向基函数时，参数为 $\sigma^2 = 1$ ， $C = 100$ 。实验结果见表 1。

由表 1 可见，当选用径向基函数作为核函数时，SVM 的分类识别率最高，并具有良好的泛化能力。除此之外，本文还将 SVM 与已有的朴素贝叶斯、BP 神经网络、kNN[11]进行比较，实验结果见表 2。

由表 2 可见，尽管线条图像特征有限，基于径向基核函数的 SVM 与其它经典方法如朴素贝叶斯、BP 神经网络、kNN 等相比仍取得更高的识别率和更好的泛化能力，因此在下述实验中，均采用基于径向基核函数的 SVM 进行实验。

为了度量所发现图像语义主题的精确度，本文按下式计算：

$$Accuracy(C2) = \frac{\sum Topic(C'_j)}{n'}, C'_j \in C2 \quad (3)$$

其中, $Topic(C'_j) = \begin{cases} 1, & \text{如果 } C'_j \text{ 的主题与实际吻合} \\ 0, & \text{否则} \end{cases}$ 。15 幅图像语义主题发现的平均精确度见表 3。

从表 3 可见, 图(1)、图(3)、图(12)、图(14)的平均精确度均在 0.6 以下, 并不理想。而平均精确度在 0.8 以上的图像有图(2)、图(4)、图(8)、图(10)、图(11)、图(13)、图(15)。仔细分析可以发现, 这些平均精确度在 0.6 以下的图像, 它们的图像主题中要么含有表示动作的词, 要么图像内容比较复杂。而平均精确度达到 0.8 以上的图像, 基本上图像内容比较单一。

为了评价语义主题发现算法的性能, 本文采用的评价指标是查准率、查全率和 $F1$ 值。在语义主题自动发现过程中, 对发现结果可分为两类: 发现成功和发现失败。发现的语义主题与实际主题不一定吻合, 可能会出现如表 4 所示的四类情形。

这里, 查全率(Recall)、查准率(Precision)、 $F1$ 值分别定义为

$$Recall = \frac{f_{11}}{f_{11} + f_{10}}, Precision = \frac{f_{00}}{f_{00} + f_{01}}, F1 = \frac{2 \times Recall \times Precision}{Recall + Precision}$$

表 5 给出了 15 幅图像的平均 $F1$ 值结果。

Table 1. Correct recognition rate of different kernel functions

表 1. 不同核函数的正确识别率

实验样本	线性核函数	多项式核函数	Sigmoid核函数	径向基核函数
训练集	84.53%	83.16%	88.06%	93.68%
测试集	81.61%	80.28%	83.75%	89.44%

Table 2. Correct recognition rate of different classifiers

表 2. 不同分类器的正确识别率

实验样本	朴素贝叶斯	BP神经网络	kNN	径向基核函数
训练集	80.68%	88.70%	90.03%	93.68%
测试集	77.16%	82.94%	85.82%	89.44%

Table 3. Average accuracy of image semantic topics

表 3. 图像语义主题的平均精确度

图(1)	图(2)	图(3)	图(4)	图(5)	图(6)	图(7)	图(8)
0.4282	0.8257	0.5304	0.8918	0.6249	0.6573	0.6970	0.8144
图(9)	图(10)	图(11)	图(12)	图(13)	图(14)	图(15)	
0.6102	0.8057	0.9016	0.5699	0.8724	0.5334	0.9126	

Table 4. Discovery results of image semantic topics on the testing set

表 4. 测试集上图像语义主题的发现结果

	成功发现主题的图像数目	不能发现主题的图像数目
实际主题的图像数目	f_{00}	f_{01}
实际非主题的图像数目	f_{10}	f_{11}

Table 5. Average F1 values of 15 images

表 5. 15 幅图像的平均 F1 值

	图(1)	图(2)	图(3)	图(4)	图(5)	图(6)	图(7)	图(8)
F1	0.1593	0.4157	0.3204	0.4516	0.3822	0.2897	0.3790	0.4439
	图(9)	图(10)	图(11)	图(12)	图(13)	图(14)	图(15)	
F1	0.3260	0.4054	0.4198	0.2021	0.3750	0.2157	0.4223	

Table 6. Discovery results of image semantic topics on the testing set using different block sizes

表 6. 不同分块尺寸下测试集上图像语义主题的发现结果

分块尺寸	图(2)	图(4)	图(8)	图(10)	图(11)	图(15)
8 × 8	0.0185	0.1127	0.1406	0.0813	0.1095	0.0895
64 × 64	0.4157	0.4516	0.4439	0.4054	0.4198	0.4223
128 × 128	0.4240	0.4289	0.4013	0.4172	0.3417	0.4317

从表 5 可见, 图(2)、图(4)、图(8)、图(10)、图(11)、图(15)的 F1 值都大于 0.4, 本算法能达到比较理想的性能。而图(3)、图(5)、图(7)、图(9)、图(13)虽然没有达到 0.4, 仍超过了 0.3, 因此, 从总体而言, 本算法的性能基本上令人满意。

为了观察分块大小对本文所提出的图像语义主题发现算法性能的影响, 本文采用的图像分块大小分别为 8 × 8、64 × 64、128 × 128。实验重复 10 次, 取平均 F1 值作为衡量性能的标准。实验结果见表 6。

由表 6 可知, 采用三种不同分块大小所对应的图像语义主题发现算法的平均 F1 值是不同的。我们发现, 8 × 8 分块的平均 F1 值最低, 基本上难以实现图像语义主题的正确发现。这是因为, 此时每个分块尺寸过小, 每块所包含的内容无法与目标语义对应。对于 64 × 64 分块, 图(4)、图(8)、图(11)取得了最高的平均 F1 值; 而对于 128 × 128 分块, 图(2)、图(10)、图(15)则取得了最高的平均 F1 值。这表明, 采用 64 × 64 分块与 128 × 128 分块时所对应的图像语义主题发现算法性能相差不多, 都能取得令人满意的结果。这是因为, 较大的分块能蕴含更丰富的语义, 但也不是越大越好, 这完全取决于所选择的图像。

8. 结论

目前, 图像语义主题的自动发现是一个研究成果较少的领域, 其研究仍有多个难点问题亟待解决。本文方法的特点在于: (1) 图像分割技术本身并不成熟, 尤其对于中国古代小说中的图像, 利用传统的图像分割方法来发现图像语义困难较大。(2) 图像分割的结果是要获得图像中的所有目标, 而本文的目的并不在此。本文关注的是图像的语义主题, 而不是图像的所有目标。(3) 已有的图像语义发现中, 颜色起很重要的作用, 而中国古代小说中的插图没有颜色特征, 这使得已有的语义发现方法不能很好地用于线条图像。(4) 线条能够很好地表达主题, 而分块更能简单明了地分析图像语义主题。实验结果也证实了本文的猜想。

从上述讨论可见, 对于图像内容较复杂并且带有动作性质的图像, 要自动发现其语义主题还有很多工作要做, 这也正是将来要继续努力的方向。

基金项目

本文系教育部人文社会科学规划基金项目(11YJAZH040)的研究成果之一。

参考文献 (References)

- [1] 张素兰, 郭平, 张继福, 胡立华 (2012) 图像语义自动标注及其粒度分析方法. *自动化学报*, 5, 688-697.

- [2] 王梅, 周向东, 等 (2008) 基于扩展生成语言模型的图像自动标注方法. *软件学报*, **9**, 2449-2460.
- [3] Jin, C. and Guo, J.L. (2014) Image semantic annotation approach based on the feature matching. *Springer Series: Advances in Intelligent Systems and Computing*, **250**, 281-288.
- [4] Carneiro, G., Chan, A.B., Moreno, P.J. and Vasconcelos, N. (2007) Supervised learning of semantic classes for image annotation and retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **29**, 394-410.
- [5] 赵琦, 张智雄, 孙坦, 许雁冬 (2009) 主题发现技术方法研究. *情报理论与实践*, **4**, 103-108.
- [6] 郭建永, 蔡勇, 甄艳霞 (2008) 基于文本聚类技术的主题发现. *计算机工程与设计*, **6**, 1426-1432.
- [7] 王小华, 徐宁, 志群 (2011) 基于共词分析的文本主题词聚类与主题发现. *情报科学*, **11**, 1621-1624.
- [8] 陈友, 程学旗, 杨森 (2011) 面向网络论坛的高质量主题发现. *软件学报*, **8**, 1621-1624.
- [9] Li, Z.D., Wang, Y., Chen, J., Xu, J. and Larid, J. (2010) Image topic discovery with saliency detection. *British Machine Vision Conference (BMVC)*, Aberystwyth, 30 August 2010-2 September 2010.
- [10] Christiani, N. and Shawe-Taylor, J. (2000) An introduction to support vector machines and other kernel based learning methods. Cambridge University Press, Cambridge.
- [11] Duda, R. and Hart, P. (1973) Pattern classification and scene analysis. John Wiley & Sons, New York.