Computer Science and Application 计算机科学与应用, 2018, 8(9), 1309-1316
Published Online September 2018 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2018.89141

Hans汉斯

Automatic Image Annotation Based on Hidden Markov Model and Convolutional Neural Network

Haijiao Xu, Qionghao Huang, Fan Wang, Yao Wen, Meihua Zhao

School of Information Technology in Education, South China Normal University, Guangzhou Guangdong
Email: guesskkk99@163.com, 893422847@qq.com, 1083202905@qq.com, 634289916@qq.com, 2454749378@qq.com

Received: Aug. 6th, 2018; accepted: Aug. 21st, 2018; published: Aug. 28th, 2018

Abstract

Automatic image annotation is becoming increasingly important in order to develop algorithms that are able to search and browse large-scale image databases. In this paper, we propose a novel annotation approach termed HMM + CNN, which is based on Hidden Markov Model (HMM) and Convolutional Neural Network (CNN). First, a multi-label CNN is trained as a concept classifier. Then, through a first-order HMM, image content and semantics correlation is combined to refine the predicted semantic scores. Finally, to improve the performance of labeling rare concepts, the gradient descent algorithm is applied for compensating the varying frequencies of concepts derived from imbalanced image datasets. Experiments have been carried out on IAPR TC-12 image annotation database. The results show that our proposed approach performs favorably compared with several conventional methods.

Keywords

Automatic Image Annotation, Hidden Markov Model, Convolutional Neural Network, Multi-Label Learning

基于隐马尔科夫模型和卷积神经网络的图像标注方法

徐海蛟，黄琼浩，汪　凡，文　瑶，赵美华

华南师范大学教育信息技术学院，广东 广州
Email: guesskkk99@163.com, 893422847@qq.com, 1083202905@qq.com, 634289916@qq.com, 2454749378@qq.com

文章引用: 徐海蛟, 黄琼浩, 汪凡, 文瑶, 赵美华. 基于隐马尔科夫模型和卷积神经网络的图像标注方法[J]. 计算机科学与应用, 2018, 8(9): 1309-1316. DOI: 10.12677/csa.2018.89141

收稿日期：2018年8月6日；录用日期：2018年8月21日；发布日期：2018年8月28日

摘　要

开发大规模图像库的搜索和浏览算法，使得图像自动标注的重要性日益增强。基于隐马尔科夫模型(HMM)与卷积神经网络(CNN)，我们提出了一种新的图像标注方法HMM + CNN。首先，训练一个多标签学习的CNN网络作为概念分类器；其次，通过一阶HMM模型把图像内容与语义相关性相结合以精炼该CNN的预测分数；最后，为改善对稀疏概念的标注性能，应用梯度下降算法来补偿在真实应用中不平衡图像集上标注概念的频率差。在IAPR TC-12标准图像标注数据集上对比了其他传统方法，结果表明我们的标注方法在查准率和查全率上性能更优。

关键词

图像自动标注，隐马尔可夫模型，卷积神经网络，多标签学习

1. 引言

随着互联网技术与多媒体共享社区的不断发展，大量的多媒体内容已进入我们的日常生活，如何高效准确地对海量的未标注图像等媒体内容进行搜索、浏览、管理变得尤为重要，这也使得图像自动标注的重要性日益增强。近年来众多学者对图像自动标注方法做了大量的研究，取得了若干阶段性成果，例如浅度学习方法：支持向量机 SVM [1]、核典型相关分析 KCCA-2PKNN [2]、稀疏核学习 SKL-CRM [3]、快速标注 FastTag [4]、离散多重伯努利模型 SVM-DMBRM [5]、图像距离尺度学习 NSIDML [6]、生成判别联合模型 GDM [7]；以及最近流行的深度学习方法：渐进式深度自动图像标注 ADA [8]、图像标签对齐模型 SEM [9]和图拉普拉斯正则化深度神经网络 HQ-III [10]等。这些传统的图像标注方法考虑了视觉特征与语义概念之间的关联，而在标注概念之间语义关联方面还存在诸多未得到很好解决的问题。很多方法仅在平衡的小概念字典上完成，而在带有大概念字典的数据集上，语义概念分布或者语义概念出现频率呈现较大差异(即概念的不平衡性)，这大大影响了标注方法的效果。因此，研究在不平衡图像库上的自动图像标注很有必要也很有意义。

在图像标注领域，深度学习方法(如卷积神经网络 CNN)比传统浅度学习方法在性能上大大提升，然而，其并未很好考虑语义概念之间的关联，这影响了其性能的进一步改善。本文针对该问题，提出了一种基于隐马尔科夫模型(HMM)与卷积神经网络(CNN)的自动图像标注方法 HMM + CNN，该方法使用 HMM 模型来校正语义标签：把图像标注过程视为检索有相互关联的隐藏语义概念序列过程，它提高了高度关联的相关概念语义分数而弱化了毫无关联的概念语义分数，提高了标注精度。在 HMM 模型里，所有的隐状态可以构成一条一阶马尔可夫链，而每个隐状态代表一个隐藏语义概念，两个隐状态之间的边权重表示它们的语义相关性，隐状态到可观测状态之间的边表示由 CNN 分类器产生的视觉语义分数。在学习过程中，考虑到真实图像集上语义概念分布的不均衡性，引入了语义概念的权重学习，其在计算发射概率和转移概率的过程中减弱了频繁概念的权重，而提升稀疏概念的权重，于是大大提高了稀疏概

念标注的性能。最后，把我们的标注方法 HMM + CNN 应用于标准标注图像集 IAPR TC-12 [11]，结果表明我们提出的标注方法 HMM + CNN 标注精度比较高，是自动图像标注的一种有效方法。

2. 隐马尔科夫模型

隐马尔可夫模型(Hidden Markov Model，简称 HMM) [12]可表达离散时间序列状态数据，它的隐状态 X_i(隐变量)不能直接观察到，但能通过观测向量 O_i 序列间接观察到。每个观测向量都是通过某些概率密度分布表现为各种状态，每一个观测向量是由一个具有相应概率密度分布的状态序列产生。所以，隐马尔可夫模型是一个双重随机过程——具有一定隐状态数的马尔可夫链和随机函数集，即两个状态集合与三个矩阵。两个状态集合是指隐状态集$\{X_1, X_2, \ldots\}$和观察状态集$\{O_1, O_2, \ldots\}$。HMM 的假设是隐状态 X_1 之间是一个马尔可夫链，对应一个初始状态矩阵 π、隐藏状态转移矩阵 $A = (a_{ij})$和发射矩阵 $B = (b_{ij})$。如图 1 所示，虚线箭头与实线箭头分别表示转移概率 a_{ij} 与发射概率 b_{ij}，当前隐状态 X_1 (隐变量)不是独立被确定出来，而是依赖于前 x 个观测向量的隐藏状态 X_{i-1}。通过使用双重随机过程，HMM 模型可以寻找到最合适的隐藏变量序列。在本模型 HMM + CNN 中，为了简化求解，仅考虑 $x = 1$ 时候的一阶情况。

假若分别使用隐语义标注 w 和未标注测试图像 I 替换 HMM 模型中的隐状态与可观测向量，则转移矩阵 A 与发射矩阵 B 分别体现了语义概念信息与视觉内容信息。我们提出的标注方法 HMM + CNN 把图像标注过程视为一个关联隐语义检索过程，与经典的标准 HMM 模型相比较，HMM + CNN 不需要传统的复杂的维特比算法，也不需要估算观测向量 I 的概率分布。

3. 基于 HMM 与 CNN 的自动图像标注方法

3.1. 问题描述

我们的标注方法 HMM + CNN 将一个图像标注过程看作是一个图像隐语义的检索过程。设训练集为 $\Delta = \{I_1, \cdots, I_{|\Delta|}\}$，其中 I_i 表示一副含有若干语义标注的图像。语义标注 w_i 来自于包含$|D|$个不同语义概念的概念字典 D，如“dance”、“horizon”和“flower”。测试集 Ω 中的图像没有包含任何语义概念标注。给定测试图像 $I \in \Omega$，图像自动语义标注(隐语义检索)的目标是在概念字典 D 中，检索出最相关的语义概念集$\{w_1, w_2, \cdots, w_K\}$，以描述 I 的视觉内容。

在检索的过程中，HMM + CNN 标注方法能够同时结合视觉内容的相关性与语义概念的相关性，在一个用户查询概念字典 D 的时候，其目标是希望检索到与未标注图像内容相一致的语义概念。一般情况

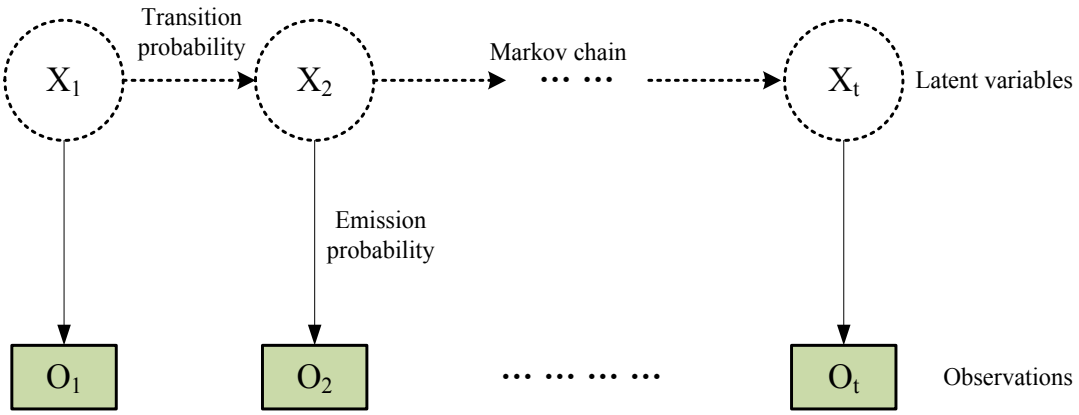

Figure 1. The graphical illustration of hidden Markov model
图 1. 隐马尔可夫模型示例图

下，用户提交的检索图像 I 在一次检索过程中是不变的，于是，未标注图像 $I \in \Omega$ 可以看做是一个可观测向量，随着检索返回语义概念数的增加，该图像 I 可重复地构成一个可观测序列$\{I(1), I(2), \cdots, I(K)\}$。

对于转移概率矩阵 $A = (a_{ij})$中的每一个元素 a_{ij} 表示两个隐状态间的转移数据，在我们的标注方法 HMM + CNN 中将 a_{ij} 视为两个语义概念 w_i 和 w_j 的关联性。而 CNN 分类器产生的视觉相关分数可看做是相应隐状态的发射概率 $b_{ij} \in B$，该发射概率 b_{ij} 表示了 CNN 分类器把图像 I 映射到语义概念 w_j 的过程。根据上述的转移概率矩阵 A 和发射矩阵 B，第 $t-1$ 步被检索的隐藏语义概念 w_i 可以关联概率值 a_{ij} 转移到第 t 步被检索隐藏语义概念 w_j，被检索隐藏语义概念 w_j 与图像 I 以发射概率值 b_{ij} 关联。由于 I 是固定不变的，所以 b_{ij} 简记为 b_j。第 t 步隐藏语义概念检索 w_j 依赖两个因素，即 $t-1$ 步检索到的隐语义概念 w_i 到该隐语义概念 w_j 的内关联转移概率 a_{ij} 以及视觉关联产生的发射概率 b_j。从上述分析可知，检索序列中相邻隐藏语义概念 w_i 与 w_j 具有相关的视觉内容与语义概念，整个相互关联的语义概念序列可以描述 I 的“故事”线索。

图 2 中给出了 HMM + CNN 标注方法的结构，矩形框表示一个未标注的检索图像 I，圆形框表示被检索的隐语义概念 w_j (隐状态)。虚线箭头表示隐状态之间的转移概率 a_{ij} 即语义关联性，而实线箭头表示每个隐藏状态的发射概率即视觉相关性 b_j。每一个隐状态 w_j 可以提供语义概念关联性数据 a_{ij} 与视觉内容相关性数据 b_j，然后基于 HMM + CNN 算法实现了对隐语义概念的排序输出。为了叙述方便，表 1 定义了 HMM + CNN 标注方法所使用的主要符号标记。

3.2. 发射概率估算

CNN 分类器产生的视觉相关性分数可视为相应隐状态的发射概率 b_j，该发射概率表示了 CNN 分类器把图像 I 映射到语义概念 w_j 的过程。任何 CNN 网络都可融入我们的标注模型，不失一般性，我们选择了近年来一个有影响力的高效 CNN 模型 ResNet [13]来作为我们的 CNN 分类器。

传统 CNN 网络聚焦于单概念分类，而我们的标注任务是一个多概念分类任务，因此为 ResNet 模型

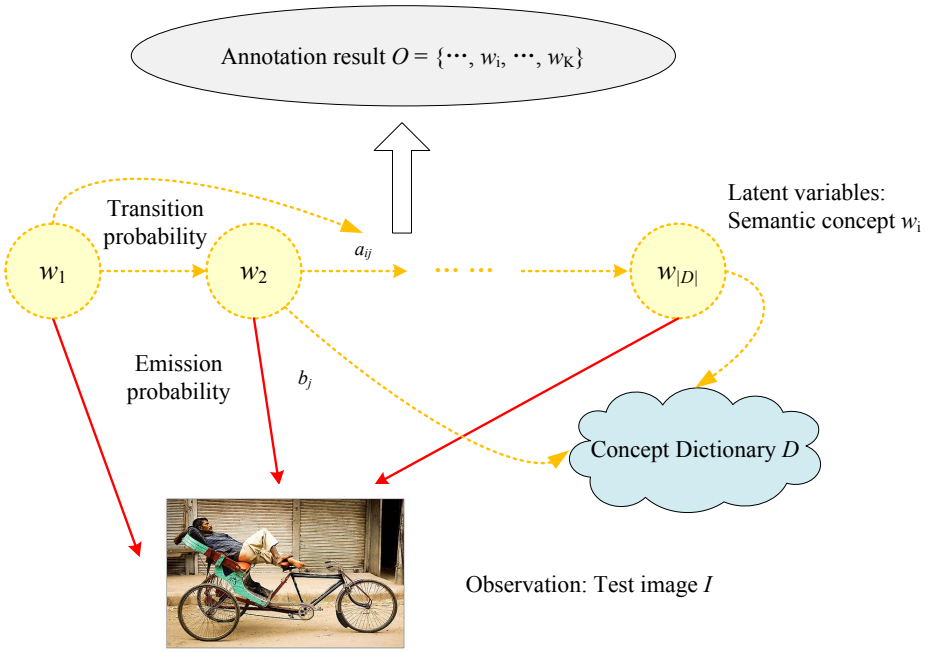

Figure 2. The HMM + CNN framework
图 2. HMM + CNN 结构图

Table 1. List of main notations
表 1. 符号标记表

符号	含义
D	概念字典
N	被检索隐语义概念集的大小，$N = \|D\|$
K	返回隐藏语义标注的数量，$K \le \|D\|$
I	被检索的未标注图像，$I \in \Omega$
$S(w)$	出现语义概念 w 的图像集合
a_{ij}	语义概念 w_i 和 w_j 之间的转移概率，即二者的语义关联概率
A	概念词典 D 的状态转移矩阵，$A = \{a_{ij} \mid i,j = 1, \cdots, N\}$
b_j	语义概念 w_j 的发射概率，即 w_j 与 I 的视觉内容相关概率
B	概念词典 D 的发射概率矩阵，$B = \{b_j \mid j = 1, \cdots, N\}$
$R(w_i)$	语义概念 w_i 的语义邻居集
K_R	语义邻居集 $R(w_i)$ 的大小，即 $K_R = \|R(w_i)\|$

定义了一个新的多概念 softmax 损失函数使之适应多概念标注任务。首先，第 i 张图片 I 与第 j 个概念 w_j 的归一化关联概率可定义为：

$$p\left(w_j \mid I\right) = \frac{\exp\left(q_j\left(I\right)\right)}{\sum_k \exp\left(q_k\left(I\right)\right)}, \tag{1}$$

其中，$q_j\left(I\right)$ 是图像 I 在第 j 个概念 w_j 的离散概率分布，它由 ResNet 分类器产生。为最小化 ResNet 预测概率与真实概率的 KL 距离，我们使用如下多概念 softmax 损失函数：

$$f_{\text{softmax}} = -\frac{1}{N}\sum_i \sum_j \bar{p}_{i,j} \log\left(p\left(w_j \mid I\right)\right), \tag{2}$$

其中，$\bar{p}_{i,j}$ 是一个图片 I 的指示器函数：当概念 w_j 在图片 I 中存在则 $\bar{p}_{i,j} = 1$ 否则 $\bar{p}_{i,j} = 0$。

3.3. 转移概率估算

两个语义概念之间的转移概率 a_{ij} 可视为二者的语义关联概率。共现概念是指以一定频率共同出现于文档中的语义概念。对于共现于图像 I 中的两个关联语义概念，由于它们共同描述了一副图像的主题或者“故事”线索，所以，可以依据共现率来估算二者的语义关联。

给定隐语义概念 w_i 和 w_j，在训练图像集Δ上考虑使用如下共现度量来计算二者的语义关联概率，HMM + CNN 标注模型以此作为两概念的转移概率 a_{ij}：

$$a_{ij} = \frac{\left|S\left(w_i\right) \cap S\left(w_j\right)\right|}{\left|S\left(w_i\right)\right|} \tag{3}$$

该公式描述了概念 w_i 和 w_j 共现的频率，然后被语义概念 w_i 的频率归一化。它可以被理解为给定含有标注 w_i 的图像 I，其包含语义概念 w_i 的概率有多大，其值范围是[0.0, 1.0]。

如果隐语义概念 w_i 是 w_j 的语义近邻，即 $w_i \in R(w_j)$，那么转移概率 a_{ij} 的值为二者的语义关联性，否则，该值被设置为 0。由于自转移的概率值很大($a_{jj} = 1$)，这导致被检索的隐语义概念可能会一直自循环在某个隐状态，导致输出无效的隐藏语义检索结果，所以自转移的概率值 a_{jj} 被设置为 0。综上所述，我们的 HMM + CNN 标注方法可以抽象为$\wedge = (N, K, I, S(w), A, B, R(w_i), K_R)$。

3.4. HMM + CNN 图像标注算法

给定一个测试集 Ω，若任意的一副图像 $I \in \Omega$ 被看做是检索对象，那么概念词典 D 中的全部语义概念 $w_j \in D$ 可构成被检索数据集。给定一个由检索对象 I 所构成的一个固定观测序列，HMM + CNN 标注方法的目标是返回能恰当描述图像 I 的最佳隐语义概念序列 $O = \{O_1, \cdots, O_j, \cdots, O_K\}$，$Oj \in D$ 的选择取决于 a_{ij} 和 b_j 两个元素，算法 1 总结了 HMM + CNN 隐语义标注的检索方法。其中，p_1 和 p_2 是待优化参数，且满足约束条件：$p_1 + p_2 = 1$，它们表示发射概率(视觉相关性)与转移概率(语义关联性)二者的权重。考虑到真实图像集的概念词典 D 是不平衡的，即不同语义概念 w 上的图像集合 $S(w)$的大小是有差异的，不同语义概念的权重p_1和p_2需要通过在训练图像集上以交叉验证方法获取，而不是直接经验设定为某个固定值。

算法 1 HMM + CNN 自动图像标注

输入：训练集Δ，概念字典 D，未标注图片 $I \in \Omega$；

输出：标注结果集 $O = \{O_1, O_2, \cdots, O_K\}$。

1 构建状态转移矩阵 $A = \{a_{ij} | i, j = 1, \cdots, N\}$；

2 构建发射概率矩阵 $B = \{b_j | j = 1, \cdots, N\}$；

3 初始化标注结果集 $O_1 = \max_{1 \le j \le N} \left(b_j\right)$, $O = O_1$；

4 for $k = 2$ to K do

5 设上一步检索概念 $O_{k-1} = w_i$，则 $O_j = \max_{1 \le j \le N} \left(p_1 b_j + p_2 a_{ij}\right)$；

6 $O = O \cup O_j$；

7 end for

8 返回结果集 O。

4. 实验和评价

4.1. 数据集

评价实验采用了公开标注数据集 IAPR TC-12 [11]。它包含有 19,627 张图像，每张图像含有 1~23 个标注，单词表 D 含有 291 个语义概念。采用随机抽样，17,665 张图像作为训练集，余下 1962 张图像作为测试集，约 75%概念频率低于平均概念频率。我们采用与文献[8]相同的评价指标：平均准确率 P、平均召回率 R、调和均值 F1 与正召回概念数 N^+。所有指标值越高表示标注性能越好。

4.2. 实验结果与分析

为观测发射概率和转移概率的语义权重$\{p_1, p_2\}$的影响，考虑使用交叉验证方法：训练图像被随机分成等份 5 组，当每组图像集交替构成验证集时，其余 4 组图像集则组成一个训练集。

首先，考虑第一种情况：忽略图像集上语义概念的不平衡性，直接给权重参数$\{p_1, p_2\}$赋经验值，观测其对于标注性能的影响。该方法记为 HMM + CNN (without weight learning)。在第二种情况，考虑语义概念分布并非是平衡的，对不同出现频次的语义概念给予相同的经验权重会导致标注性能的下降，因此，对于不同的语义概念 $w_i \in D$ 给予不同权重$\{p_1, p_2\}$，全部概念权重组成不同的发射概率向量 P1 与转移概率向量 P2，{P1, P2}可在验证集上用如下方法求出来。

1) 初始化权重向量，即 P1 = 0, P2 = 1 − P1；

2) 对于任意的 $i(1 \le I \le N)$：

2a) 对于不同的权重 $p_1 \in \{0, 0.1, \cdots, 1\}$执行算法 1 获得标注结果集 O 并记录最大 F1 性能分数，写入相对应的权重值 pm_i：$P1_i = pm_i$, $P2_i = 1 - P1_i$；

Table 2. Performance comparisons of multi-label image annotation
表 2. 多标签图像标注性能比较

标注方法	平均准确率 P	平均召回率 R	调和均值 F1	正召回概念数 N+
SVM [1]	0.27	0.31	0.29	157
KCCA-2PKNN [2]	0.59	0.30	0.39	259
SKL-CRM [3]	0.47	0.32	0.38	274
FastTag [4]	0.47	0.26	0.34	280
SVM-DMBRM [5]	0.56	0.29	0.38	283
NSIDML [6]	0.57	0.37	0.45	282
GDM [7]	0.32	0.29	0.30	252
ADA [8]	0.42	0.30	0.35	280
SEM [9]	0.41	0.39	0.40	-
HQ-III [10]	0.43	0.41	0.42	281
HMM + CNN (Ours)	**0.64**	**0.45**	**0.53**	**285**

2b) i 值加一，即 $i = i + 1$；

3) 输出权重向量 P1, P2。

显然，上述权重提升方法的时间复杂度是 O(11 × N)。从实验结果见，第二种情况下的基于权重学习方法的 HMM + CNN 图像标注方法效果优于第一种情况下的 HMM + CNN (without weight learning)标注方法。

表 2 列出了与最新图像标注方法的对比实验结果。

从表 2 中可见，我们的 HMM + CNN 方法超越了其他对比方法，获得了更好的标注性能。与表中最好的对比标注方法 NSIDML 比较，HMM + CNN 方法的平均准确率、平均召回率、调和均值 F1 分别提高了 12%、22%、18%。一方面，视觉内容的相关性(发射概率)可以挖掘视觉内容与语义概念的相关性，另一方面，语义关联性(转移概率)反映出隐标注概念之间的语义关联，其更准确地描述了图像的“故事”线索。在图像标注任务中，这两种类型的相关性都提供了有用的信息，具有一定的互补性，从这个角度上说我们的 HMM + CNN 方法可提高图像标注的性能。此外，在隐语义标注检索中使用了语义权重学习方法来获得合理的语义权重，所以，在不平衡数据集上，我们的 HMM + CNN 方法具有更好的标注效果。

参考文献

[1] Chang, C.C. and Lin, C.J. (2011) LIBSVM: A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology*, **2**, 1-27. https://doi.org/10.1145/1961189.1961199

[2] Ballan, L., Uricchio, T., Seidenari, L. and Bimbo, A.D. (2014) A Cross-Media Model for Automatic Image Annotation. *International Conference on Multimedia Retrieval*, 73. https://doi.org/10.1145/2578726.2578728

[3] Moran, S. and Lavrenko, V. (2014) Sparse Kernel Learning for Image Annotation. *International Conference on Multimedia Retrieval*, 113-120. https://doi.org/10.1145/2578726.2578734

[4] Chen, M., Zheng, A. and Weinberger, K. (2013) Fast Image Tagging. *International Conference on Machine Learning*, 1274-1282.

[5] Murthy, V.N., Can, E.F. and Manmatha, R. (2014) A Hybrid Model for Automatic Image Annotation. *International Conference on Multimedia Retrieval*, 369-376. https://doi.org/10.1145/2578726.2578774

[6] Jin, C. and Jin, S.W. (2016) Image Distance Metric Learning Based on Neighborhood Sets for Automatic Image Annotation. *Journal of Visual Communication and Image Representation*, **34**, 167-175.

https://doi.org/10.1016/j.jvcir.2015.10.017

[7] Ji, P., Gao, X. and Hu, X. (2017) Automatic Image Annotation By Combining Generative and Discriminant Models. *Neurocomputing*, **236**, 48-55. https://doi.org/10.1016/j.neucom.2016.09.108

[8] 周铭柯, 柯逍, 杜明智. 基于数据均衡的增进式深度自动图像标注[J]. 软件学报, 2017, 28(7): 1862-1880.

[9] Ma, Y., Liu, Y., Xie, Q. and Li, L. (2018) CNN-Feature Based Automatic Image Annotation Method. *Multimedia Tools & Applications*, 1-14. https://doi.org/10.1007/s11042-018-6038-x

[10] Mojoo, J., Kurosawa, K. and Kurita, T. (2017) Deep CNN with Graph Laplacian Regularization for Multi-Label Image Annotation. *International Conference on Image Analysis and Recognition*, 19-26. https://doi.org/10.1007/978-3-319-59876-5_3

[11] Grubinger, M., Clough, P. and Müller, H. (2006) The IAPR Benchmark : A New Evaluation Resource for Visual Information Systems. *International Conference on Language Resources and Evaluation*, 13-23.

[12] Saini, R., Roy, P. and Dogra, D. (2018) A Segmental HMM Based Trajectory Classification Using Genetic Algorithm. *Expert System Application*, **93**, 169-181. https://doi.org/10.1016/j.eswa.2017.10.021

[13] He, K., Zhang, X., Ren, S. and Sun, J. (2016) Deep Residual Learning for Image Recognition. *IEEE Conference on Computer Vision and Pattern Recognition*, 770-778.