一种语义级文本协同图像识别方法

段喜萍[1,2,3],刘家锋[1],王建华[2,3],唐降龙[1]

(1.哈尔滨工业大学 计算机科学与技术学院,150001 哈尔滨; 2. 哈尔滨师范大学 计算机科学与信息工程学院,150025 哈尔滨; 3. 黑龙江省高校智能教育与信息工程重点实验室, 150025 哈尔滨)

摘 要: 为解决单纯依赖图像低级视觉模态信息进行图像识别准率低的问题. 考虑到许多图像中存在文本信息,提出了利用图像中的文本信息辅助图像识别的语义级文本协同图像识别方法. 该方法通过文本定位方法定位到图像中的文本块,对其进行分割、二值化、提取特征等处理;然后获取语义,提取图像底层视觉信息,计算两模态的相关性,从而得到协同后验概率; 最后,得到联合后验概率,并取其中最大联合后验概率对图像进行识别. 在自建体育视频帧数据库中,通过与以朴素贝叶斯为代表的单模态方法进行比较,方法在3种不同视觉特征下均具有更高的准确率. 实验结果表明,文本协同方法能够有效辅助图像识别,具有更好的识别性能.

关键词: 文本定位;图像识别;多模态

中图分类号: TP391. 41 **文献标志码**: A **文章编号**: 0367-6234(2014)03-0049-05

A collaborative image recognition method based on semantic level of text

DUAN Xiping[1,2,3], LIU Jiafeng[1], WANG Jianhua[2,3], TANG Xianglong[1]

(1.School of Computer Science and Technology, Harbin Institute of Technology, 150001 Harbin, China;
2.Computer Science and Information Engineering College, Harbin Normal University, 150025 Harbin, China;
3. Heilongjiang Provincial Key Laboratory of Intelligence Education and Information Engineering, 150025 Harbin, China)

Abstract: To solve the problem that singular-modal image recognition using only the low-level visual features has low accuracy, considering that many images have embedded-in textual information, a collaborative method using the embedded-in text to aid the recognition of images is proposed. The method includes three steps. Firstly, after localization, segmentation, binarization and feature extraction, semantics of text is gotten. Secondly, the collaborative posterior probability is calculated by extracting visual features of images and counting correlation of visual and textual modals. At last, for each class of images, the joint posterior probability is calculated using the previous two items. A new image is recognized to the class with maximal joint posterior probability. Experiments on the self-built data set of sports video frames showed that the proposed method performed better than the singular-modal method on three different visual features and had higher accuracy.

Keywords: text localization; image recognition; multi-modal

如何有效地对图像或视频等多媒体信息进行分类和识别,以实现诸如图像自动标注、图像检索等应用具有重要意义,也是目前一个迫切需要解决的热点问题.在图像识别中,由于“语义鸿沟”的存在,单纯利用图像底层视觉信息往往不能达到很好的识别效果.同时许多图像中包含着与图像语义更为相关的文字或文本信息,并且这种图像的数目相当可观,如图1所示.而从目前的情况来看,对于这类图像,存在不同角度的研究,关心图像内容识别的一般不关心其中包含的文本信息,将其视作与问题无关的背景或者是场景的一部分来处理;而关心图像文本识别的则在检测出文本所在的区域之后就不再关心图像其他部分的内容

收稿日期: 2013-05-24.
基金项目: 国家自然科学基金资助项目 (61173087,41071262).
作者简介: 段喜萍(1980—),女,博士研究生;
唐降龙(1960—),男,教授,博士生导师.
通信作者: 段喜萍,xpduan_1999@126.com.

了. 不论是图像识别还是文本识别都没有完整地利用图像中所包含的视觉和文本两种模态信息进行识别,图像中的信息被孤立地处理. 而其中一种模态信息对另一种态的语义识别具有重要意义. 例如,图1给出的一组建筑物图像,借助图像中的文本信息,很容易对它们进行区分和识别. 着眼于此,本文研究利用图像上的文本信息辅助图像内容识别.

图1 包含嵌入文本的图像

据进行文献搜索所掌握的资料来看,目前还没有同时利用图像视觉信息以及其上的文本信息进行图像识别的先例. 与之相关的研究有:1)基于图像底层信息进行图像识别,即基于计算机视觉的图像识别,该类方法可进一步分为判别式方法[1-3]和产生式方法[4-11]. 由于"语义鸿沟"现象的存在,不能保证视觉特征相似的图像在语义上也相近. 因而该类方法无法实现对图像内容的准确识别. 2)对图像场景文本进行检测与识别[12-13]. 该类方法在图像中检测文本区域,然后提取文本区域的字符前景,使用字符识别技术识别图像区域中的文本,一旦检测出文本所在的区域之后就不再关心图像其他部分的内容. 3)利用图像周边文本辅助图像识别[14]. 这类方法利用图像周围文本,如图像的标题、链接、锚定文本以及替代文本等,建立图像和文本之间的关联关系,辅助图像识别,这类方法适用于具有周围文本的网络图像识别.

本文提出一种能够同时利用图像视觉信息与图像上嵌入的文本信息的方法,将每个模态的识别结果作为一种最简单的语义信息用于协同,而不涉及更高层级的语义内容. 具体来说,同时提取图像视觉特征信息和文本特征信息,获取文本语义信息,然后利用文本语义信息辅助图像视觉信息进行建模,建立联合后验概率. 模型可分解为:单模态文本语义识别、单模态图像内容识别以及两模态类别相关程度计算. 通过对以上模型的训练,建立各图像类识别器,对新图像进行识别.

1 协同模型

利用文本模态辅助图像视觉模态进行图像识别的过程可以看作是一种利用"跨模态(cross-modality)"信息进行识别的过程. 单模态的识别过程一般是在观察到属性特征 x 的条件下对类别属性 ω 的后验概率 $P(\omega|x)$ 进行建模的过程. 而在跨模态假设之下,其中某一模态类别属性的后验概率需要使用两个模态的特征属性共同建模. 即对图像类别 ω_I 的识别不仅需要图像视觉模态的特征 x_I,同时还需要考虑图像中文本模态的特征 x_T,即需要对 $P(\omega_I|x_I,x_T)$ 进行建模,这里将 $P(\omega_I|x_I,x_T)$ 称作联合后验概率,它可以通过以下两种方式建模.

1.1 多模态直接建模

从理论上讲,联合后验概率只是扩大了识别对象的特征属性集合,可以采用一般的识别方法进行建模,即通过扩大特征向量维数直接对多模态信息建模. 然而对于实际问题来说,直接对联合后验概率建模往往存在着一定的困难,原因是:

1)特征的描述方式不同. 来自于不同模态的特征可能是以不同方式描述的,如图像内容特征可以用颜色或梯度直方图描述,显著性区域的散列表示,甚至是采用多示例包的方式描述;而文本和文字特征则可以描述为笔划的密度,傅里叶变换、小波变换系数、笔划之间的结构关系等等. 按照不同方式描述的特征很难采用统一的形式建模,更适合于分别采用不同的模型描述.

2)模型学习困难. 即使来自于不同模态的特征可以采用相同的方式描述,如果将两个模态的特征组合为扩大的特征集合,势必造成描述联合后验概率的模型的复杂度的增加. 而在图像识别的实际应用中,可获得的学习样本一般是有限的,采用数量不足的样本学习一个复杂的模型,无法保证模型的泛化能力.

1.2 多模态协同建模

为解决联合后验概率直接建模和学习的困难,本文提出使用文本模态辅助视觉模态对联合后验概率建模,如图2所示.

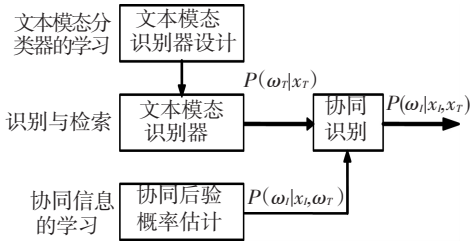


图2 语义级文本协同的图像识别过程

模型主要通过 Stieltjes 积分实现联合后验概率的简化,具体简化为

$$P(\omega_I|x_I,x_T)\approx\int P(\omega_I|x_I,\omega_T)\mathrm{d}F(\omega_T|x_T). \quad (1)$$

式中:$P(\omega_I|x_I,\omega_T)$ 为文本模态对视觉模态的语义级协同后验概率估计;$P(\omega_T|x_T)$ 对应于分布

函数 $F(\omega_T \mid x_T)$，表示文本模态识别器. 从上述模型可以看出，语义级协同模型的关键是协同后验概率 $P(\omega_I \mid x_I, \omega_T)$ 的估计. 协同后验概率也可以采用两种方式计算：直接方式和间接方式.

1）直接方式. 已知某模态特征和另一个模态语义类别信息条件下，协同后验概率描述的是该模态类别的发生概率. 因此可以在学习阶段，按照一个模态的类别监督信息将样本划分成不同的子集，分别学习当该模态属于某个类别时另一个模态的分类器. 协同识别时根据一个模态的识别结果分别使用不同的分类器计算另一个模态的类别后验概率.

直接方式计算的好处是可以模型化一个模态的特征与另一个模态语义类别信息之间的关联性，其缺点是学习时需要对样本集合进行划分，这就造成了单个分类器的学习样本较少，降低了模型的泛化能力.

2）间接方式. 一般情况下，假设一个模态的特征与另一个模态的语义信息之间相互独立是合理的，例如在不同场景下，某字符的特征是由所属文本类别决定的，与其所处的环境无关. 在此假设下，协同后验概率可被简化为

$$P(\omega_I \mid x_I, \omega_T) = P(\omega_I \mid x_I) \frac{P(\omega_I, \omega_T)}{P(\omega_I) P(\omega_T)}. \tag{2}$$

将式（1）、（2）结合可以得到

$$P(\omega_I \mid x_I, x_T) = \int P(\omega_I \mid x_I) P(\omega_T \mid x_T) \frac{P(\omega_I, \omega_T)}{P(\omega_I) P(\omega_T)} \mathrm{d}\omega_T. \tag{3}$$

由于语义类别信息是一个离散随机变量，因此协同识别可以将式（1）和式（3）的 Stieltjes 积分转化为有限求和式直接进行计算. 这样式（3）可转化为

$$P(\omega_I \mid x_I, x_T) = \sum_{\omega_T = 1}^{c} P(\omega_I \mid x_I) P(\omega_T \mid x_T) \cdot \frac{P(\omega_I, \omega_T)}{P(\omega_I) P(\omega_T)}. \tag{4}$$

式中：c 为文本语义类别数. 需要强调一点，上述模型适合于仅利用图像中的单字协同图像识别的情况. 考虑到多字情况，如 N 个字，则可对式（1）~（4）进行扩展，得到

$$P(\omega_I \mid x_I, x_{T_1}, x_{T_2}, \cdots, x_{T_N}) = \sum_{\omega_{T_1}=1}^{c} \cdots \sum_{\omega_{T_N}=1}^{c} P(\omega_I \mid x_I) \prod_{i=1}^{N} P(\omega_{T_i} \mid x_{T_i}) \frac{P(\omega_I, \omega_{T_i})}{P(\omega_I) P(\omega_{T_i})}. \tag{5}$$

2　模型计算与学习

由式（5）可以看到，该模型可归结为 3 部分：$P(\omega_T \mid x_T)$、$P(\omega_I \mid x_I)$ 以及 $\frac{P(\omega_T, \omega_I)}{P(\omega_T) P(\omega_I)}$，分别表示单模态的文本识别、图像内容识别以及两模态类别之间的相关程度计算. 在学习过程中，可对各文本模态类别和视觉模态类别分别进行学习，并对两模态类别之间的相关程度进行统计. 然后在测试阶段，可将测试集样本分别代入该模型（即式（5））计算各图像类别的联合后验概率，并按照最大联合后验概率对各测试样本进行分类.

2.1　文本识别

文本识别的过程可归结为文本定位、分割、二值化、特征提取以及识别的过程. 其中前两个过程本文采用文献[12]的方法，首先将图像划分成块，通过滤波器结合边分析进行文本定位；然后对确定的文本块分别进行垂直和水平投影，通过得到的垂直和水平柱条进行文本分割. 对分割出的每个字符图像进行二值化处理后可将得到的二进制文本块放缩到某一指定大小，并拉成一行向量，经 PCA 处理后得到最终的文本向量，即特征向量. 对通过以上过程得到的一组训练样本，采用朴素贝叶斯方法可得到各文本类结构 $P(\omega_T \mid x_T)$，其中：$\omega_T = 1, 2, \cdots, c_T$，$c_T$ 为文本类别总数.

2.2　图像内容识别

图像内容识别可以根据具体应用提取相应的视觉特征，构造相应的识别器. 当需要对整体场景属性分类时，可以以颜色分布、纹理特征为基础构建图像分类器，而当需要识别图像中某类目标时，则需要提取图像的局部描述特征（如显著性区域，Blob 区域特征等）构成 Bag of Features，然后采用 Constellation 模型或多示例的方式构造分类器. 本文在实验中分别提取了颜色分布特征、小波纹理特征以及 Blob 量化特征. 同样，在识别器设计过程中，基于提取的图像视觉特征采用朴素贝叶斯方法训练出多个视觉模态识别器结构 $P(\omega_I \mid x_I)$，其中：$\omega_I = 1, 2, \cdots, c_I$，$c_I$ 为图像类别总数.

2.3　两模态类别相关程度计算

两模态类别间的相关程度，可以在学习过程中根据样本的监督信息统计得到. 具体来说，需要估计两模态语义信息的同现概率 $P(\omega_T, \omega_I)$ 以及每个模态类别的先验概率 $P(\omega_T)$ 和 $P(\omega_I)$，然后通

过 $\frac{P(\omega_T,\omega_I)}{P(\omega_T)P(\omega_I)}$ 计算两模态类别间的相关程度. 该相关程度可在有限样本条件下较为准确地估计.

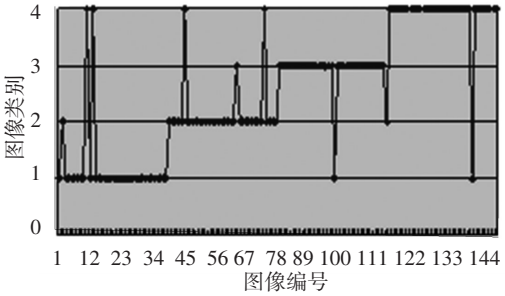
3 实验设置与结果分析

3.1 数据集和实验设置

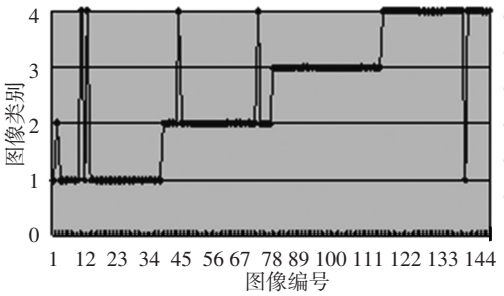
为了检验本文本协同模型对图像识别问题的有效性,这里对本协同模型与单模态分类器的识别性能进行比较. 单模态分类器选择了朴素贝叶斯方法.

本实验采集的图像数据是从 CCTV5 网络电视台下载的包括篮球、排球、短道速滑、羽毛球等 4 类视频,从中抽取了 360 张带文本的视频帧,建立体育视频图像数据库,其中 3/5 用于训练,2/5 用于测试,即有 216 张用于训练,144 张用于测试. 对文本数据,在训练阶段,一部分取自前面带文本的图像中的文本,另一部分为人工生成文本. 增加人工生成文本的目的,主要是扩大文本样本数量,提高识别的泛化能力;在测试阶段的文本,直接取自从测试图像中获得的文本. 需要注意的是,本文在视频图像中提取的文本主要是体育视频在后期制作中所添加的标题文本(caption text 或 superimposed text),做这样的选择主要基于两个原因:1)标题文本是人为添加的,与视频内容具有更强的相关性和概括性;2)标题文本相对于可能出现的场景文本(scene text)更清晰、更容易识别,并且识别准确率高. 本文在标题文本定位过程中,除了使用文献[12]中的方法外,还考虑到标题文本通常在位置、高、宽等方面的限制,从而大大排除了场景文本的影响.

本文在实验过程中,分别提取了图像的全局颜色、全局纹理、Blob 特征,设定的文本类别为 8 类,其目的是验证本文本协同方法是否对不同的图像视觉特征具有普遍适用性.

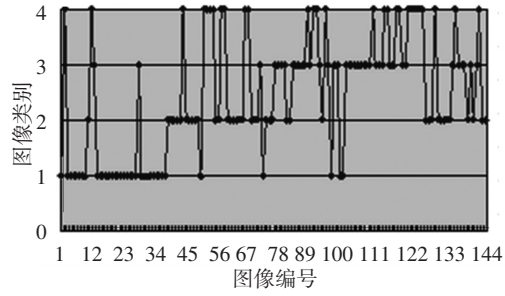


(a)单模态分类结果

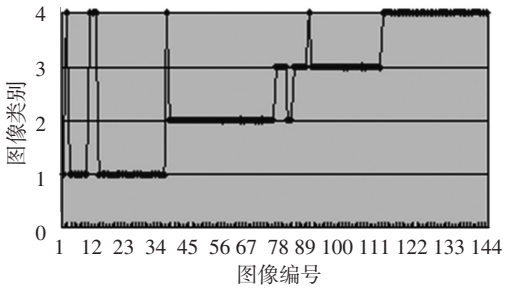


(b)本文模型分类结果

图 3 视觉特征取全局颜色特征的分类结果

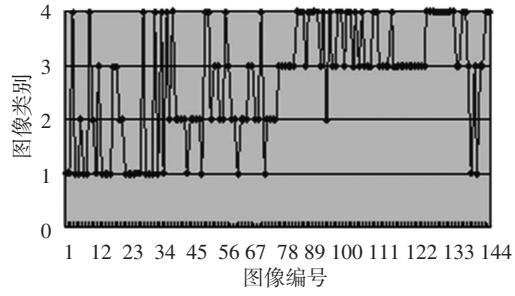


(a)单模态分类结果

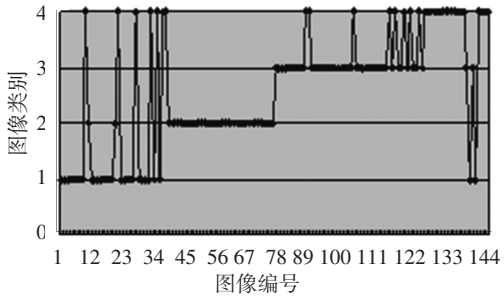


(b)本文模型分类结果

图 4 视觉特征取全局纹理特征的分类结果



(a)单模态分类结果



(b)本文模型分类结果

图 5 视觉特征取 Blob 特征的分类结果

3.2　结果比较与分析

在与单模态方法进行的对比实验中，图像识别的性能通过识别准确率进行衡量，具体的识别结果如图 3～5 所示，准确百分率如表 1 所示. 实验表明：由于本文方法使用了除视觉模态信息外的文本模态信息，图像表示更为全面和恰当. 因此，本文方法的识别结果优于单模态方法. 特别是在单模态识别性能较低的情况下，本文方法能显著提高性能. 如在视觉特征取全局纹理时，单模态方法的准确率为 64.58%，本文方法的准确率可达 94.44%，提高了 29.86%；在视觉特征取 Blob 特征时，单模态方法的准确率为 53.47%，本文方法的准确率可达 82.64%，提高了 29.17%.

表 1　识别准确率比较　　%

方　法	全局颜色	全局纹理	局部斑块
单模态方法	93.75	64.58	53.47
文本协同方法	95.83	94.44	82.64

对上述结果进行分析，可以得出：对选择的不同图像视觉特征，本文的文本协同方法都一定程度地提高了图像识别的准确率，从而验证了本文最初想法的正确性和合理性. 需要指出的是，本方法取得较好效果取决于文本识别要有较高精度，因而对标题文本进行操作可得到理想结果.

4　结　论

1）本文提出一种使用文本模态信息辅助图像视觉模态进行图像识别的方法，即一种语义级文本协同图像识别方法. 其优势在于：能够全面地利用图像中的视觉模态信息和文本模态信息.

2）该图像识别方法的精度取决于选择的图像特征、选择的文本类别数以及文本分类器的分类能力等方面. 在选择的几种图像视觉特征上实验，本识别方法的识别准确率均高于单模态方法.

3）需要指出的是，本文选择的文本是体育视频图像中相对清晰的标题文本，对场景文本情况并不理想.

参考文献

[1] PANDA N, CHANG E Y. Efficient top-k hyperplane query processing for multimedia information retrieval [C]//Proceedings of the 14th annual ACM international conference on Multimedia. New York, NY: ACM, 2006: 317-326.

[2] LU Zhiwu, IP H H S. Image categorization with spatial mismatch kernels[C]//IEEE Conference on Computer Vision and Pattern Recognition. Miami, FL: IEEE, 2009: 397-404.

[3] SONG X, JIAO L C, YANG S, et al. Sparse coding and classifier ensemble based multi-instance learning for image categorization [J]. Signal Processing, 2013, 93(1):1-11.

[4] RUSSELL B C, FREEMAN W T, EFROS A A, et al. Using multiple segmentations to discover objects and their extent in image collections[C]//IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2006: 1605-1614.

[5] VAILAYA A, FIGUEIREDO M A T, JAIN A K, et al. Image classification for content-based indexing [J]. IEEE Transactions on Image Processing, 2001, 10(1): 117-130.

[6] LI F F, PERONA P. A bayesian hierarchical model for learning natural scene categories[C]//IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2005: 524-531.

[7] LIU D, CHEN T. Unsupervised image categorization and object localization using topic models and correspondences between images [C]//International Conference on Computer Vision. Piscataway. NJ: IEEE, 2007: 1-7.

[8] FERGUS R, PERONA P, ZISSERMAN A. Object class recognition by unsupervised scale-invariant learning[C]// IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2003: 264-271.

[9] LIU Y, GOTO S, IKENAGA T. A robust algorithm for text detection in color images[C]//Proceedings of the Eighth International Conference on Document Analysis and Recognition. Piscataway, NJ: IEEE, 2005: 399-403.

[10] CHEN Y, WANG J Z. Image categorization by learning and reasoning with regions[J]. The Journal of Machine Learning Research, 2004, 5(12): 913-939.

[11] ZHU L, ZHAO B, GAO Y. Multi-class multi-instance learning for lung cancer image classification based on bag feature selection [C]//Fifth International Conference on Fuzzy Systems and Knowledge Discovery. Piscataway, NJ: IEEE, 2008: 487-492.

[12] SHIVAKUMARA P, HUANG W, TAN C L. An efficient edge based technique for text detection in video frames [C]//The Eighth IAPR International Workshop on Document Analysis Systems. Piscataway, NJ: IEEE, 2008: 307-314.

[13] MISHRA A, ALAHARI K, JAWAHAR C V. Top-down and bottom-up cues for scene text recognition[C]//2012 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2008: 2687-2694.

[14] 许红涛，周向东，向宇，等. 一种自适应的 Web 图像语义自动标注方法[J]. 软件学报，2010，21(9)：2183-2195.

（编辑　张　红）