文章编号:1672-3961(2009)01-0022-05

基于协同训练的指纹图像分割算法

周广通,尹义龙*,郭文鹃,任春晓
(山东大学计算机科学与技术学院, 山东 济南 250101)

摘要:指纹图像分割是自动指纹识别系统预处理中最关键的技术之一.精确、可靠地将指纹图像从背景中分割出来,能够加快后续工作的处理速度,提高识别算法的准确性.传统的分割算法需要大量已标记的指纹图像作为训练数据,但实际应用中获取标记样本比较繁琐和耗时.为综合利用已标记和未标记的指纹图像,提出一种基于协同训练的半监督指纹图像分割算法:CoSeg.该算法在基于像素水平的 Coherence、Mean、Variace(CMV)特征体系下,使用标记盒和支持向量机作为基分类器进行协同训练.在 FVC2002 指纹库上的实验结果表明,CoSeg 能够在标记信息较少的情况下取得较好的性能,并在处理低质量指纹图像时表现出较强的鲁棒性.

关键词:指纹识别;指纹图像分割;半监督学习;协同训练;CoSeg
中图分类号:TP391;TP18 **文献标志码**:A

Fingerprint segmentation algorithm based on co-training

ZHOU Guang-tong, YIN Yi-long*, GUO Wen-juan, REN Chuan-xiao
(School of Computer Science and Technology, Shandong University, Jinan 250101, China)

Abstract: Fingerprint segmentation is one of the key preprocessing steps in an automated fingerprint identification system (AFIS). Effective segmentation of the fingerprint from the background could speed up following processes and improve recognition accuracy. However, in traditional segmentation algorithms, it is essential to obtain lots of labeled fingerprints, which are usually more expensive than unlabeled ones. To incorporate labeled and unlabeled data together, this paper proposed CoSeg, a semi-supervised fingerprint segmentation algorithm. Under the view of pixel-level features, i.e. Coherence, Mean and Variance (CMV), CoSeg employs Label Box and SVM as two base learners and trains the final model for segmentation based on a co-training style algorithm. Experiments performed on FVC 2002 databases show that CoSeg can effectively exploit unlabeled data with limited labeled data, and the proposed method is also robust when dealing with low-quality fingerprints.

Key words: fingerprint recognition; fingerprint segmentation; semi-supervised learning; co-training; co-training based fingerprint segmentation algorithm

0 前言

指纹图像分割是自动指纹识别系统(automated fingerprint identification system, AFIS)预处理中的一个重要环节[1].精确可靠地分割不仅能够减少后续工作的处理时间,而且可以显著改进特征提取的可信度,提高识别的准确性.指纹图像分割本质上是一个两类分类问题,即将图像分为前景区和背景区,仅前景区含有指纹,需要保留以做后续处理.

根据特征提取对象的不同,指纹图像分割算法可分为基于块水平的和基于像素水平的2种.基于

收稿日期:2009-01-19
基金项目:山东省优秀中青年科学家科研奖励基金资助项目(2006BS01008);山东省高新技术自主创新工程专项资助项目(2007ZCB01030);山东省自然科学基金重点项目资助项目(Z2008G05)
作者简介:周广通(1986-),男,山东菏泽人,硕士研究生,研究方向为机器学习、数据挖掘与模式识别.
E-mail:zhouguangtong@gmail.com
* **通讯作者**:尹义龙(1972-),男,山东菏泽人,教授,博士,研究方向为图像处理、模式识别与机器学习.
E-mail:ylyin@sdu.edu.cn

块水平的算法,通常把把指纹图像分成若干块图像,然后对每个块提取特征并作为一个数据样本使用,如文献[2]等,该类算法处理速度较快但分割精度有限;基于像素水平的算法则对像素提取特征并形成数据样本,如文献[3-5],该类算法能够得到较为平滑的分割边界但相对比较耗时.本文工作在基于像素水平的指纹图像分割算法上展开.

目前主流的指纹图像分割算法多采用监督学习方式,需要大量已标记(这里的标记是指某一像素是前景点或背景点)的数据进行训练以使模型具有良好的泛化能力.但在实际应用中,获取已标记指纹图像通常需要用人工方式实现,比较繁琐和耗时,而获取未标记指纹图像则相当容易.因此,在保持原有分割精度的同时,如何有效地降低所需标记的数据量是指纹图像分割中一个值得研究的问题.

传统的机器学习方法,充足的监督样本可以使模型具有良好的泛化能力,但在许多实际应用中,获取大量已标记的样本则相对较为困难,因为获得这些标记可能需要付出昂贵的代价,而收集大量未标记的样本已相当容易.显然,如果只使用少量的已标记样本,那么由此训练出的分类器往往很难具有强泛化能力.另外,如果仅使用少量高代价的已标记样本而不利用大量易获取的未标记样本,也是对数据的一种浪费.因此,能够综合利用已标记样本和未标记样本的半监督学习逐渐成为机器学习领域一个新的研究热点[6].

协同训练(co-training)算法是一种典型的半监督学习方法.标准的协同训练算法是由 A. Blum 等人[7]在 1998 年提出的,他们假设样本集有 2 个充分并且冗余的视图,基于这 2 个视图可以训练得到 2 个分类器;然后,在协同训练过程中,每个分类器从未标记样本中挑选出若干置信度较高的样本进行标记,并把标记后的样本加入另一个分类器的已标记训练集中,以方便对方利用这些新标记的样本进行学习.此过程不断迭代进行,直到满足终止条件.之后,W. Wang 等人[8]从理论上证明了即使使用单个视图,只要 2 个基分类器的初始学习效果有一定差异,协同训练算法也同样有效.

本文提出了一种基于单视图协同训练的指纹图像分割算法(co-training based fingerprint segmentation algorithm,CoSeg).该算法在基于像素水平的 CMV 特征体系下,以标记盒算法[5](Label Box)和支持向量机算法[9](support vector machine,SVM)作为基分类器进行协同训练,训练完成后结合二者的输出对指纹图像中的像素进行分类,从而实现分割.CoSeg 综合利用了已标记数据和未标记数据,实验结果也表明,该算法能够在标记信息较少的情况下取得较好的性能,并在处理低质量指纹图像时有较强的鲁棒性.

1　研究背景

1.1　基于像素水平的指纹分割

A.M. Bazen 等人[3]在 2001 年提出了一种基于像素水平的指纹图像分割算法,他们首次提出了基于像素的 CMV 特征体系,即方向一致性(coherence),灰度均值(mean)和灰度方差(variance)3 个特征,并利用线性分类器对指纹图像进行分割;Y. Yin 等人[4]对该方法进行了扩展,他们使用神经网络训练得到一个二次曲面分类器,实验结果证明了该方法的有效性;任春晓等人[5]提出了一种基于标记盒(Label Box)的分割算法,该算法本质上是一个分段线性分类器,它首先把 CMV 特征空间分成若干子盒,根据落入子盒中的训练样本把每个子盒标记为前景盒或背景盒,并由该标记对像素分类.

1.2　协同训练算法

虽然 A. Blum 等人[7]在提出协同训练算法时给出了严格的理论推导来证明该算法的有效性,但在实际应用中,充分并且冗余的 2 个视图的假设往往很难得到满足.S. A. Goldman 等人[10]提出了一种不需要满足上述假设的协同训练算法,他们使用不同的决策树算法,在同一个属性集上训练出 2 个不同的分类器,每个分类器都可以把样本空间划分为若干个等价类;Z. H. Zhou 等人[11]提出了一种既不要求充分冗余假设、也不要求使用不同类型分类器的 tri-training 算法,该算法使用 3 个分类器,结合集成学习的优势来提高泛化能力.随后,W. Wang 等人[8]从理论上证明了即使使用单视图,只要 2 个基分类器的初始学习效果有较大差异,协同训练算法也能同样有效,这也为协同训练算法在充分冗余假设不满足条件下的应用提供了理论基础.目前,协同训练算法已在很多领域得到成功应用,如电子邮件分类[12]、文本分类[13]等;此外,Z. H. Zhou 等人[14]也提出了一种基于协同训练的半监督回归算法.

2　CoSeg 算法

2.1　算法流程

CoSeg 算法是基于单视图 CMV 的协同训练算法.该算法首先对已标记的指纹图像像素赋予相应的类别标记,形成已标记样本集,未标记样本集由指

纹库中的部分指纹图像得来.随后,CoSeg 采用标记盒算法与支持向量机算法作为基分类器协同训练,训练结束后得到决策模型,并根据此模型对指纹进行分割(详见 2.2 节).CoSeg 算法的具体执行流程见图 1.为控制训练时已标记样本和未标记样本的数量,算法采用 bootstrap 方法抽样.

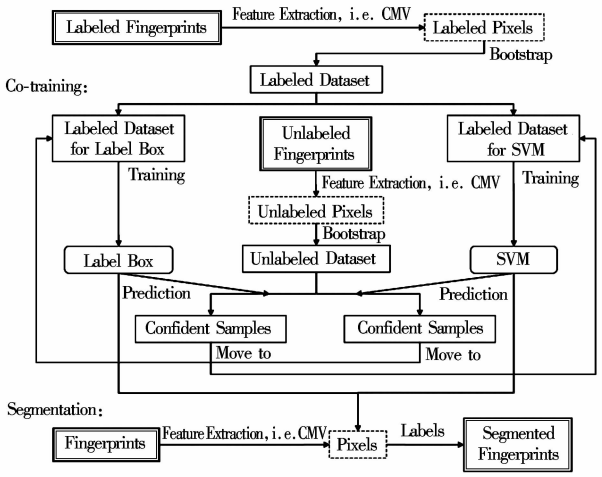

图 1 CoSeg 算法整体流程

Fig.1 Framework of CoSeg

2.2 协同训练算法

CoSeg 首先利用给定的已标记样本集训练得到初始 Label Box 模型和 SVM 模型;然后分别根据这两个分类器选择未标记样本中具有较高置信度的样本,并加入对方的训练样本集中;随后由新的训练样本集(此时 2 个训练样本集已存在一定差异)重新训练基分类器,此过程重复执行数次,可得到协同训练后的 Label Box 模型和 SVM 模型.CoSeg 使用平均后验概率的方法对 2 个基分类器集成,得到分割模型.上述协同训练算法的具体细节如算法 1 所示.

算法 1 CoSeg(P^l, P^u, k, N_{fore}, N_{back})

$P^l_{\text{LB}} \leftarrow P^l$;

$P^l_{\text{SVM}} \leftarrow P^l$;

for $i = 0: k$ do

$H^i_{\text{LB}} \leftarrow \text{LabelBox}(P^l_{\text{LB}})$;

$H^i_{\text{SVM}} \leftarrow \text{SVM}(P^l_{\text{SVM}})$;

Conf $P^i_{\text{LB}} \rightarrow$ Select N_{fore} most confidence foreground Pixels and N_{back} most confidence background pixels from P^u by H^i_{SVM};

Conf $P^i_{\text{SVM}} \leftarrow$ Select N_{fore} most confidence foreground pixels and N_{back} most confidence background pixels from P^u by H^i_{SVM};

$P^l_{\text{SVM}} \leftarrow P^l_{\text{SVM}} \cup \text{Conf}\ P^i_{\text{LB}}$;

$P^l_{\text{LB}} \leftarrow P^l_{\text{LB}} \cup \text{Conf}\ P^i_{\text{SVM}}$;

$P^u = P^u - \text{Conf}\ P^i_{\text{LB}} \cup \text{Conf}\ P^i_{\text{SVM}}$;

end for

return $H_{\text{CoSeg}} = (H^k_{\text{LB}} + H^k_{\text{LB}})/2$;

算法 1 的输入中,P^l 表示已标记的像素点集,P^u 表示未标记的像素点集,k 为迭代次数,N^c_{fore} 和 N^c_{back} 分别为每次加入对方训练样本集的前景像素点和背景像素点数.训练时,初始得到的 Label Box 模型和 SVM 模型分别为 H^0_{LB} 和 H^0_{SVM};经过 k 次迭代的协同训练后,用平均后验概率的方法集成 Label Box 模型和 SVM 模型得到 CoSeg 模型 H_{CoSeg}.分割时,CoSeg 首先对分割指纹图像 P^s 的所有像素点预测类别标记,即对 $p \in P^s$:如果 $\text{Prob}_{\text{CoSeg}}(p) \geqslant 0.5$,则预测 p 为前景点(foreground),否则预测 p 为背景点(background),其中 $\text{Prob}_{\text{CoSeg}}$ 可由 H_{CoSeg} 计算得来.所有像素点预测完成后即可按相应的标记对指纹图像分割.

3 实验

3.1 实验设置

为验证 CoSeg 算法的有效性,本文实验使用 FVC 2002 指纹库中 db1_a、db2_b 和 db3_b 3 个数据集作为测试对象,并同 Label Box 算法和 SVM 算法做对比.其中,Label Box 算法中子盒个数设置为 $50 \times 50 \times 50$;SVM 算法基于 LIBSVM[15] 实现,其核函数采用线性核函数,CoSeg 算法中基分类器的设置同上,迭代次数设为 10 次,每次迭代 2 个基分类器分别标记2 000 个前景点和 1 000 个背景点并加入到对方的训练样本集中.

实验首先选取 5 副指纹图像进行人工标记,并分别从中抽样出 1×10^1, 1×10^2, 1×10^3, 1×10^4 和 1×10^5 个像素点形成已标记样本集;然后随机选取 5 幅指纹图像并抽样出 1×10^5 个像素点形成未标记样本集.Label Box 模型和 SVM 模型由各个已标记样本集训练得到,而 CoSeg 模型则由各个已标记样本集和未标记样本集分别结合形成的训练样本集训练得到.此外,在每个数据集(b1_a、db2_b 和 db3_b)上分别另选 13 幅指纹图像进行人工标记并形成测试数据集.上述训练和预测过程重复 3 次,最终的实验结果是由这 3 次的结果平均得到.实验在 Pentium (R)D 2.80 GHz,512 M 内存的计算机上进行.

本文选用算法错误率(error rate, ER)和算法预测时间(time cost, TC)来评价分割性能,其中 ER 的计算为

$$R_{\text{E}} = \frac{N^c_{\text{back:fore}} + N^c_{\text{fore:back}}}{N^c_{\text{back}} + N^c_{\text{fore}}},$$

其中 N^c_{back} 和 N^c_{fore} 分别为人工分割结果中前景和背景像素点的个数,$N^c_{\text{back:fore}}$ 和 $N^c_{\text{fore:back}}$ 分别为算法中错

分的前景和背景的像素点个数.

3.2 实验结果及分析

算法错误率的统计结果如表 1 ~ 3 所示.

表 1　FVC 2002 db1_a 测试数据集上 ER 对比

Table 1　Comparison of ER on FVC 2002 db1_a database　%

已标记样本数量	R_E		
	Label Box	SVM	CoSeg
1×10^1	7.522 8	7.677 7	6.472 6
1×10^2	6.653 2	5.856 8	5.016 5
1×10^3	4.494 2	4.364 5	4.494 2
1×10^4	4.063 6	3.633 0	4.054 7
1×10^5	3.717 3	3.278 6	3.711 0

表 2　FVC 2002 db2_b 测试数据集上 ER 对比

Table 2　Comparison of ER on FVC 2002 db2_b database　%

已标记样本数量	R_E		
	Label Box	SVM	CoSeg
1×10^1	13.690 3	13.029 5	12.734 9
1×10^2	10.530 3	8.331 6	11.723 0
1×10^3	12.152 1	12.526 3	14.438 4
1×10^4	13.575 6	16.304 0	12.504 5
1×10^5	13.767 0	19.255 4	12.485 0

表 3　FVC 2002 db3_b 测试数据集上 ER 对比

Table 3　Comparison of ER on FVC 2002 db3_b database　%

已标记样本数量	R_E		
	Label Box	SVM	CoSeg
1×10^1	11.870 4	9.558 8	2.897 8
1×10^2	10.419 3	9.382 7	2.820 8
1×10^3	11.544 0	17.335 4	6.287 5
1×10^4	26.056 7	12.452 2	8.706 4
1×10^5	12.027 1	28.435 7	9.496 0

由算法错误率统计结果可知,在已标记样本数较少时,CoSeg 算法的性能优于 Label Box 和 SVM,这说明 CoSeg 能够在标记信息较少的情况下充分利用未标记的信息,而当标记样本数增多时,由于可利用的标记信息足够支持算法建模,故 CoSeg 与 Label Box 和 SVM 算法的性能也逐步趋于一致.指纹图像的质量对算法性能的影响是至关重要的,如在质量较好的数据集上(如 db1_a 和 db2_b),CoSeg、Label Box 和 SVM 的性能相当;但在质量较差的数据集(如 db3_b)上,CoSeg 体现出较强的鲁棒性,这是因为基分类器在处理低质量指纹图像时差异性较大,而 CoSeg 的协同训练过程能够有效地集成基分类器的结果,最终取得一个较好的分割性能.此外,在质量较差的数据集(如 db3_b)上,可能由于 bootstrap 随机抽样引入了较多的噪声点,CoSeg、LabelBox 和 SVM 的分割错误率随着标记样本的增多反而呈现上升的趋势,标记样本的质量对分割效果的影响也是十分关键的.

算法预测时间的统计结果如表 4 所示,该结果为每个测试数据集上 CoSeg、LabelBox 和 SVM 算法预测时间的平均.

表 4　算法预测时间对比

Table 4　Comparison of time cost　s

数据集	t_C		
	Label Box	SVM	CoSeg
db1_a	0.040 7	0.020 3	0.053 2
db2_b	0.048 7	0.025 4	0.055 7
db3_b	0.025 1	0.010 6	0.026 4

由于预测时需要集成已训练完成的 Label Box 模型和 SVM 模型,故 CoSeg 算法的预测时间要高于 LabelBox 算法和 SVM 算法,尽管如此,这个预测时间还是可以接受的.

图 2 和图 3 给出了部分典型图像及相应的 CoSeg 算法分割结果,图中的分割结果均未做后续处理.

图 2　db1_a 1_5.tif 分割效果(从左至右分别为原始图像、Label Box 分割效果、SVM 分割效果和 CoSeg 分割效果(已标记样本数为 100))

Fig.2　Segmentation of db1_a 1_5.tif(from left to right: original fingerprint, segmented image by Label Box, segmented image by SVM, segmented image by CoSeg)

图 3　db1_a 110_1.tif 分割效果(从左至右分别为原始图像、Label Box 分割效果、SVM 分割效果和 CoSeg 分割效果(已标记样本数为 100))

Fig.3　Segmentation of db1_a 110_1.tif(from left to right: original fingerprint, segmented image by Label Box, segmented image by SVM, segmented image by CoSeg)

4　结语

本文提出的 CoSeg 算法在基于像素水平的 CMV 特征体系下,以 Lable Box 和 SVM 作为基分类器进行协同训练,并有效地利用了已标记数据和未标记数据.实验结果表明:CoSeg 能在标记信息较少的情况下取得较好的分割效果,并可在处理质量较差的指

纹图像时表现出较强的鲁棒性.

基于 2 个独立视图的 CoSeg 是后续工作的主要方向,如频域视图和空域视图,像素视图和块视图等;自动选取有代表性的已标记或未标记像素点,而不是采用抽样的方式,也是值得关注的研究内容,此外,不同基分类器和不同参数(如 k, N^C_{fore} 和 N^C_{back} 等)对分割效果的影响也有待于进一步研究.

参考文献:

[1] JAIN A K, PANKANTI S, PRABHAKAR S, et al. Recent advances in fingerprint verification[C]// Proceedings of the 3rd Audio- and Video-Based Biometric Person Authentication (AVBPA'01). Halmstad, Sweden: Springer, 2001: 182-191.

[2] CHEN Xinjian, TIAN Jie, CHENG Jiangang, et al. Segmentation of fingerprint images using linear classifier[J]. EURASIP Journal on Applied Signal Processing, 2004, 2004(1): 480-494.

[3] BAZEN A M, GEREZ S H. Segmentation of fingerprint images [C]// Processing of the ProRISC 2001, Workshop on Circuits, Systems and Signal Processing (ProRISC'01), Netherlands: IEEE, 2001: 276-280.

[4] YIN Yilong, WANG Yanrong, YANG Xiukun. Fingerprint image segmentation based on quadric surface model[C]// Proceedings of the 5rd Audio- and Video-Based Biometric Person Authentication (AVBPA'05). Hilton Rye Town, NY: Springer, 2005: 647-655.

[5] 任春晓,尹义龙. 基于标记盒的指纹分割[J]. 山东大学学报:工学版, 2006, 36(5):54-57.
REN Chunxiao, YIN Yilong. Fingerprint image segmentation based on label box[J]. Journal of Shandong University: Engineering Science, 2006, 36(5):54-57.

[6] ZHU Xiaojin. Semi-supervised learning literature survey[R]. Technical report 1530, Madison, WI: Department of Computer Sciences, University of Wisconsin, 2007.

[7] BLUM A, TOM M Mitchell. Combining labeled and unlabeled data with co-training[C]// Proceedings of the 11th Annual Conference on Computational Learning Theory (COLT'98), Madison, WI: ACM, 1998: 92-100.

[8] WANG W, ZHOU Zhi-Hua. Analyzing co-training style algorithms[C]// Proceedings of the 18th European Conference on Machine Learning (ECML'07). Warsaw, Poland: Springer, 2007: 454-465.

[9] VAPNIK V N. The nature of statistical learning theory[M]. New York: Springer, 1995.

[10] GOLDMAN S A, ZHOU Yan. Enhancing supervised learning with unlabeled data[C]// Proceedings of the 7th International Conference on Machine Learning (ICML'00), Standord: Morgan Kaufmann, CA, 2000: 327-334.

[11] ZHOU Zhi-Hua, LI Ming . Tri-training: exploiting unlabeled data using three classifiers[J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(11):1529-1541.

[12] KIRITCHENKO Svetlana, MATWIN Stan. Email classi-fication with co-training[C]// Proceedings of the 2001 conference of the Centre for Advanced Studies on Collaborative Research (CASCON'01), Toronto, Canada: IBM, 2001: 8.

[13] NIGAM Kamal, GHANI Rayid. Analyzing the effectiveness and applicability of co-training[C]// Proceedings of the 17th Conference on Information and Knowledge Management (CIKM'00), McLean, VA: ACM, 2000: 86-93.

[14] ZHOU Zhi-Hua, LI Ming. Semisupervised regression with co-training-style algorithms[J]. IEEE Transactions on Knowledge and Data Engineering, 2007, 19(11):1479-1493.

[15] CHANG Chihchung, LIN Chihjen. LIBSVM: a library for support vector machines[EB/OL]. 2001. Software available. [2009-01-15]. http://www.csie.ntu.edu.tw/~cjlin/libsvm.

(编辑:孙培芹)