

基于图的高维数据分类 Ratio Cut 模型及其快速算法

郑世秀^{1,2} 潘振宽¹ 徐知磊¹

(青岛大学计算机科学技术学院 山东 青岛 266071)¹ (青岛大学商学院 山东 青岛 266071)²

摘 要 数据分类是数据挖掘研究的重要内容,随着数据量以及数据维度的增加,对大规模、高维数据的处理成为关键问题。为提高数据分类的准确率,受计算机视觉中图像分割算法的启发,针对经典的 Ratio Cut 分类模型提出一种基于非局部算子的实现算法。引进拉格朗日乘子,建立新的能量泛函,并采用交替优化的策略来求解该能量泛函。数值实验表明,算法的准确率及计算效率与传统分类方法相比都有较大提高。

关键词 图,非局部方法,Ratio Cut,数据分类

中图法分类号 TP391 文献标识码 A

Graph-based Ratio Cut Model for Classification of High-dimensional Data and Fast Algorithm

ZHENG Shi-xiu^{1,2} PAN Zhen-kuan¹ XU Zhi-lei¹

(College of Computer Science and Technology, Qingdao University, Qingdao, Shandong 266071, China)¹

(Business School, Qingdao University, Qingdao, Shandong 266071, China)²

Abstract Data classification is an important part of data mining. With the increase of the amount of data and the dimension of data, the processing of large-scale and high-dimensional data becomes the key problem. In order to improve the accuracy of data classification, inspired by the image segmentation algorithm in computer vision, an algorithm based on nonlocal operator was proposed for the classic Ratio Cut classification model. A new energy functional is modeled by introducing Lagrange multipliers, and the energy functional is solved by the alternating optimization method. Numerical experiments show that the accuracy and computational efficiency of the proposed algorithm are greatly improved compared with the traditional classification method.

Keywords Graph, Nonlocal means, Ratio Cut, Data classification

1 引言

信息时代的飞速发展带来了数据量的爆炸式增长。随着数据量的增大,数据分类问题越来越困难。数据分类方式主要有两种,一种是无监督分类,另一种是监督或半监督分类。为有效解决大数据中的分类问题,迫切需要高效的数据分类方法。在众多的分类方法中,常用的有支持向量机、神经网络、深度学习以及基于图的分类聚类方法等。这些方法各有优势,其中支持向量机比较适合小样本,深度学习适合大样本,而基于图的分类聚类方法比较直观^[1]。在这些方法中,基于图的分类方法近年来在半监督学习中应用得非常广泛。在全监督学习如支持向量机和深度学习中,往往需要大量的训练样本,而在半监督学习中,只需要少量的样本就可以将数据进行归类,而且归类好的数据可以直接用来训练,这样大大地提升了机器学习或深度学习的效率,从而节省大量的成本。随着研究的深入,基于图的变分分类方法不仅可以解决两类分类问题,还可以解决多分类问题,在无监督以及半监督分类和学习中具有明显优势,其主要体现在能取得较高的分类正确率。在半监督分类和学习中,需要极少的样本点就可以实现大规模的数据分类,而且分类后的数据可以直接用于机器

学习和深度学习的训练。

变分法的飞速发展源于 90 年代初, Rudin 等提出了 TV (Total Variation)^[2] 模型,并将这个模型成功地应用到图像去噪问题上,这一应用在变分法的发展中具有里程碑式的意义。变分方法被广泛应用于求解反问题,在之后的 20 多年中,变分方法被应用到图像处理的各个方面,包括图像恢复、图像分割、图像配准、三维重建以及光流计算等,而且在每个方面都有很多经典的模型。

从经典 TV 模型开始,为了简化 TV 模型的计算,研究者们提出了 Dual 算法以及 Split Bregman^[3] 算法,这些算法的提出都具有非常重要的意义。随后研究发现 TV 模型具有较好的边缘保持能力,但去噪后的图像会具有阶梯效应,为了克服阶梯效应,研究者们提出了很多高阶模型,使得处理后的图像变得更加平滑。随着研究的深入,为了保持图像的纹理信息, Gilboa 等^[4-5] 提出了非局部算子即 Nonlocal-TV 模型,这个模型的优势在于不仅可以保持边缘信息,让处理后的图像保持光滑,还可以保持纹理图像的纹理。Nonlocal-TV 的提出加速了变分图像处理的发展,为图像处理中其他方面的研究奠定了坚实的基础。在变分图像分割的发展过程中,最开始提出了两类分类的 Mumford-Shah^[6] 模型,随后又提出了 Chan-

本文受国家自然科学基金(61170106)资助。

郑世秀(1972—),女,博士生,讲师,主要研究方向为数字图像处理、模式识别, E-mail: 13697677910@163.com; 潘振宽(1966—),男,博士,教授,主要研究方向为数字图像处理、机器学习; 徐知磊(1992—),男,硕士生,主要研究方向为数字图像处理、模式识别。

Vese^[7]模型,这两个模型在变分图像分割领域均具有重要的意义。

随着变分图像处理方法的发展,研究者开始利用上述两个模型进行多相图像分割,在这个过程中,学者提出了用 n 个标记函数分 n 类的方法,潘振宽等^[8]提出了用 $n-1$ 个标记函数的分 n 类方法。其他方法还有用 m 个标记函数分 2^m 类的方法,这是一种非常高效的分类分割标记方法,可以极大地提升计算效率;还有用一个标记函数分 n 类的方法,这是一种更加高效的方法。这些标记方法对于变分图像分割的发展具有重要意义。为了分割纹理图像,研究者将非局部方法移植到图像分割模型,并取得了很多重要的成果。将非局部算子应用于纹理图像分割为将非局部算子应用于数据分类打下了基础。端金鸣等^[9]对非局部方法的应用做了重要探索。

在基于图的非监督学习中,随着变分方法的发展,非局部方法被移植到图上,这种方法的主要思想是基于数据的邻域,利用邻域信息将数据进行归类,从而达到数据分类的目的。基于图上非局部算子的分类方法可以结合变分图像分割方法,将图像分割中的标记函数法应用到图上的分类工作中,可以大大提高计算的效率。到目前为止,变分法在数据分类中的应用是变分法的一个最新应用,通过构造能量泛函、求解能量泛函实现数据分类。由于数据分类与图像分割有很多相似之处,因此数据分类实际上就是图上的分割问题,这使得图像分割中已有的方法有可能被应用于数据分类,这种移植不是盲目的,而是有严谨的理论研究。目前在变分领域,数据分类主要是基于非局部算子,这个非局部算子与图像处理中的不同,它是基于图的。

基于图的 Ratio Cut^[10],Cheeger Cut^[11]以及 Normalized Cut 等^[12]模型与数据分类问题密切相关。Szlam 等^[13]提出求解 Cheeger Cut 问题的松弛方法,采用与图像处理中的能量方法相似的求解方法,提出相应的 split-Bregman 算法。Bresson 等^[14]提出两种求解 Cheeger Cut 松弛问题的算法:1)新型的最速下降方法;2)一种修改的反能量方法。Bresson 等^[15]提出了运用自适应结束条件的另一个版本,其结果更加有效。这些方法都是无监督学习方法,没有任何已知的信息。

基于图的多分类方法也取得了一定的进展。一种方法是使用二分类方法的递归求解,另一种是设计专门的多分类求解算法。Bresson 等^[16]提出不使用递归的多类聚类方法,采用近似的 split-Bregman 算法求解。随后 Bresson 等^[17]进行了扩展,在已知一小部分数据点的标记的前提下同时处理多类。Hu 等^[18]研究了最小化包含总变差项的能量函数以获得模块最优化的方法。

本文基于图的相关概念引进非局部算子,采用变分方法对经典 Ratio Cut 模型重新设计其能量泛函,并为其设计了快速求解的交替优化算法。

2 图上的非局部算子

在图像分割中,一个标记函数可以将图像划分为两部分。类似地,两类数据分类问题即把数据 V 划分为 V_1 和 V_2 ,其标记函数 $u(x)$ 设计为:

$$u(x) \in \{0, 1\}, x \in V, u(x) = \begin{cases} 1, & \text{if } x \in V_1 \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

图上的数据分类问题为离散问题,因此需要将连续的非局部算子转化为离散形式。连续空间非局部算子为:

$$\partial_x u(x) = (u(y) - u(x)) \sqrt{w(x, y)}, y \in N(x) \quad (2)$$

离散空间非局部算子为:

$$\partial_x u(x) = (u(y) - u(x)) \sqrt{w(x, y)}, y \in N(x) \quad (3)$$

非局部矢量 $\mathbf{v} = v(x, y)$ 的离散非局部散度定义为:

$$(\nabla_w \mathbf{v})(x) := \sum_{y \in N(x)} (v(x, y) - v(y, x)) \sqrt{w(x, y)} \quad (4)$$

矢量 v_1 和矢量 v_2 的内积为:

$$\langle v_1, v_2 \rangle(x) := \sum_{y \in N(x)} v_1(x, y) v_2(x, y) \quad (5)$$

基于以上定义,离散化的非局部梯度模 $|\nabla_w u|(x)$ 的计算公式如下:

$$|\nabla_w u|(x) = \sqrt{\sum_{y \in N(x)} (u(y) - u(x))^2 w(x, y)} \quad (6)$$

离散非局部拉普拉斯算子 Laplacian $div_w(\nabla_w u)(x)$ 的计算公式如下:

$$div_w(\nabla_w u)(x) = 2 \sum_{y \in N(x)} (u(y) - u(x)) w(x, y) \quad (7)$$

因此,最小割问题 $MinC(V_1, V_2)$ 等价于

$$MinC(V_1, V_2) = \min_{\substack{x \in V_1 \\ y \in V_2}} \sum_{x \in V} w(x, y) = \min_{u \in \{0, 1\}, x \in V} \sum_{x \in V} |\nabla_w u|^2(x) \quad (8)$$

所表达的凸松弛优化问题为基于 Tikhonov 规则项的能量泛函极小值问题:

$$\min_{u \in [0, 1]} \{E(u) = \sum_{x \in V} |\nabla_w u|^2(x)\} \quad (9)$$

式(9)的确可以求得图上的数据分类问题的解,但在很多情况下,容易造成不平衡的分类,因此实际分类中往往采用带有约束的模型,以获得平衡分类。

3 基于图的 Ratio Cut 分类模型及其快速算法

Ratio Cut 问题即:

$$RatioCut(V_1, V_2) = cut(V_1, V_2) \left(\frac{1}{V_1} + \frac{1}{V_2} \right) \quad (10)$$

该模型以类顶点的数目作为约束,使得分类均衡。但该模型为 NP-Hard 问题,无法直接求解。下面设计其快速求解算法。

引进非局部算子,将模型(10)改写为如下能量泛函:

$$E(u) = \sum_{x \in V} |\nabla_w u|(x) \left(\frac{1}{\sum_{x \in V} u(x)} + \frac{1}{n - \sum_{x \in V} u(x)} \right) \quad (11)$$

s. t. $u \in \{0, 1\}$

其中, n 是数据总量。式(1)很难直接求解,因此采用交替优化策略和拉格朗日乘子法(Alternating Direction Method of Multipliers, ADMM)来求解这个问题,得到能量泛函如下:

$$(u^{k+1}, v^{k+1}) = \arg \min_{u, v} \left\{ \sum_{x \in V} |\mathbf{v}| \left(\frac{1}{\sum_{x \in V} u(x)} + \frac{1}{n - \sum_{x \in V} u(x)} \right) + \lambda_k \sum_{x \in V} (\mathbf{v} - \nabla_w u) + \frac{\mu}{2} \sum_{x \in V} (\mathbf{v} - \nabla_w u)^2 \right\} \quad (12)$$

其中, λ_k 是拉格朗日乘子, μ 是惩罚参数。关于 u 的欧-拉格朗日方程由下式给出:

$$g|\mathbf{v}| + \nabla_w \cdot \lambda_k + \mu \nabla_w \cdot (\mathbf{v} - \nabla_w u) = 0 \quad (13)$$

$$g = 1 / (n - \sum_{x \in V} u^k(x))^2 - 2 / (\sum_{x \in V} u^k(x))^2 \quad (14)$$

通过下式计算 u :

$$u^{k+1}(x) = \frac{\sum_{y \in N(x)} u^k(y) w(x, y) - \nabla_w \cdot \lambda_k - \mu \nabla_w \cdot (\mathbf{v}^k) - g|\mathbf{v}^k|}{2\mu \sum_{y \in N(x)} w(x, y)} \quad (15)$$

s. t. $u \in [0, 1]$

关于 v 的欧-拉格朗日方程由下式给出:

$$\frac{\mathbf{v}}{|\mathbf{v}|} \left(\frac{1}{\sum_{x \in V} u(x)} + \frac{1}{n - \sum_{x \in V} u(x)} \right) + \mu(\mathbf{v} - \nabla_w u - \frac{\lambda_k}{\mu}) = 0 \quad (16)$$

为了计算 v , 为其设计软阈值公式如下:

$$v = \max\left(\left|\nabla_w u^{k+1} - \frac{\lambda_k}{\mu}\right| - \frac{f}{\mu}, \frac{\nabla_w u^{k+1} - \frac{\lambda_k}{\mu}}{\left|\nabla_w u^{k+1} - \frac{\lambda_k}{\mu}\right|}\right) \quad (17)$$

$$0 \frac{0}{|0|} = 0$$

$$f = \frac{1}{\sum_{x \in V} u^k(x)} + \frac{1}{n - \sum_{x \in V} u^k(x)} \quad (18)$$

接下来计算 λ_k :

$$\lambda_k^{k+1} = \lambda_k^k + \mu(v^{k+1} - \nabla_w u^{k+1}) \quad (19)$$

按照以上步骤计算结束后得到 $u \in [0, 1]$, 因此最后需要将其投影到 $u \in \{0, 1\}$ 。算法描述如算法 1 所示。

算法 1 Ratio Cut-ADMM 算法(RC-ADMM)

任意初始化 u^0

$\lambda_k^0 \leftarrow 0$

while 不收敛 do

 利用式(15)计算 u^{k+1}

 利用式(17)计算 v^{k+1}

 利用式(19)计算 λ_k^{k+1}

end while

4 数值实验

本节用两个数据集验证所提出的算法, 迭代控制条件为:

$$\left| \frac{E(k) - E(k-1)}{E(k)} \right| \leq \epsilon, k \geq 2 \quad (20)$$

其中, $E(k)$ 为第 k 步的能量, $E(k-1)$ 是第 $k-1$ 步的能量, ϵ 是一个较小的值, 用来控制迭代次数。

双月数据集为交叉形数据集, 常用来测试分类算法。按照如下方法生成双月人工数据集: 在 $0 \sim \pi$ 之间取 1000 个值, 组成 1000×2 的矩阵。

$$T_1 = [\cos(x), \sin(x)]$$

$$T_2 = [\cos(x) + 1, 0.5 - \sin(x)]$$

连接两个矩阵, 扩展为 100 维, 并添加方差 $\sigma = 0.02$ 的随机噪声, 这样得到了 2000×2 的高维数据。如图 1 所示, 上下半月各由 1000 个点组成。

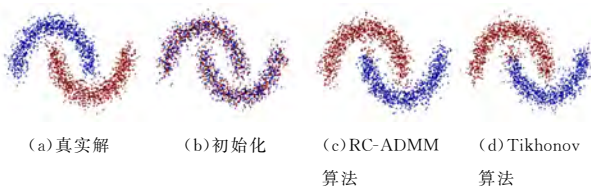


图 1 Ratio Cut ADMM 算法在双月数据集上的结果

应用本文提出 Ratio Cut ADMM 算法将双月数据集分成两类, 分别为 1030 和 970, 错误率为 1.7%。而使用 Tikhonov 算法分成的两类为 1062 和 938, 错误率为 4.2%。准确率比较如表 1 所列。

表 1 两种算法在双月数据集上的准确率比较

(单位: %)

算法	准确率
RC-ADMM 算法	98.3
Tikhonov 算法	95.8

另外两个实验采用手写体识别 MNIST 数据集。MNIST 数据集下载自 <http://yann.lecun.com/exdb/mnist/>。该数据

集由 70000 幅 28×28 的手写数字 0~9 的图像组成, 分为 10 个类别。因为数据集较大, 本文使用其中一个子集。由于数字 4 和数字 9、数字 3 和数字 8 在手写体中有时比较容易混淆, 因此分别使用由数字 3 和 8 组成的 13966 个数字和由数字 4 和数字 9 组成的 13782 个数字作为识别数据集, 使用本文的算法对其进行分类并将分类结果可视化。

数字 3 和 8 数据集的分类结果如图 2 所示。

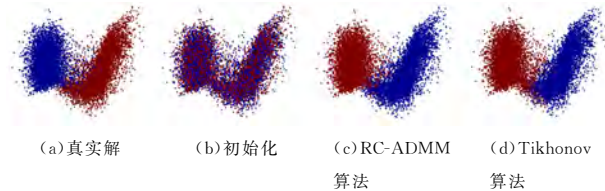


图 2 Ratio Cut ADMM 算法在数字 3 和 8 数据集上的结果

应用本文提出 Ratio Cut ADMM 算法将数字 3 和 8 数据集分成两类, 分别为 7173 个数字和 6739 个数字, 错误率为 1.24%。使用 Tikhonov 算法分成两类分别为 7320 个和 6654 个, 错误率为 2.44%。准确率比较如表 2 所列。

表 2 两种算法在数字 3 和 8 数据集上的准确率比较

(单位: %)

算法	准确率
RC-ADMM 算法	98.76
Tikhonov 算法	97.56

应用本文提出 Ratio Cut ADMM 算法将数字 4 和 9 分成两类, 分别为 6836 个和 6946 个, 错误率为 1.27%。用 Tikhonov 算法分成的两类数据为 7012 个和 6700 个, 错误率为 2.41%。其结果如图 3 所示, 准确率比较如表 3 所列。

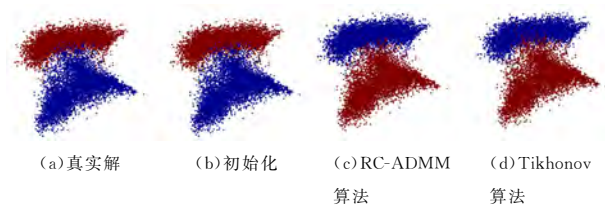


图 3 Ratio Cut ADMM 算法在数字 4 和 9 数据集上的结果

表 3 两种算法在数字 4 和 9 数据集上的准确率比较

(单位: %)

算法	准确率
RC-ADMM 算法	98.73
Tikhonov 算法	97.59

上述实验表明, 所设计的快速算法在分类准确率方面比传统方法有较大提高。经过多次实验表明, 所提算法稳定性的较好, 所得百分比均为多次实验取得的平均值。

结束语 受图像分割算法的启发, 本文提出一种利用非局部算子实现均衡分类的 Ratio Cut 模型, 并为其设计了快速求解算法——ADMM 算法。数值实验表明, 所提算法的准确率及计算效率与传统分类方法相比都有较大提高。后续研究需要对不同约束条件下的分类效果做进一步的分析, 并对高维数据的多分类算法做进一步研究。

参 考 文 献

- [1] VON LUXBURG U. A tutorial on spectral clustering[J]. Statistics and Computing, 2007, 17: 395-416.

- [2] RUDIN L I, OSHER S, FATEMI E. Nonlinear total variation based noise removal algorithms[J]. *Physica D: Nonlinear Phenomena*, 1992, 60(1): 259-268.
- [3] GOLDSTEIN T, OSHER S. The Split Bregman Method for L1-Regularized Problems[J]. *Siam Journal on Imaging Sciences*, 2009, 2(2): 323-343.
- [4] GILBOA G, OSHER S. Nonlocal Linear Image Regularization and Supervised Segmentation[J]. *Siam Journal on Multiscale Modeling & Simulation*, 2007, 6(2): 595-630.
- [5] GILBOA G, OSHER S. Nonlocal Operators with Applications to Image Processing[J]. *Siam Journal on Multiscale Modeling & Simulation*, 2008, 7(3): 1005-1028.
- [6] VESE L A, CHAN T F. A Multiphase Level Set Framework for Image Segmentation Using the Mumford and Shah Model[J]. *International Journal of Computer Vision*, 2002, 50(3): 271-293.
- [7] CHAN T, VESE L. An Active Contour Model without Edges [M]. *Scale-Space Theories in Computer Vision*. Springer Berlin Heidelberg, 1999: 141-151.
- [8] 潘振宽, 李华, 魏伟波, 等. 三维图像多相分割的变分水平集方法[J]. *计算机学报*, 2009, 32(12): 2464-2474.
- [9] DUAN J, PAN Z, LIU W, et al. Color Texture Image Inpainting Using the Non Local CTV Model[J]. *Journal of Signal and Information Processing*, 2013, 4: 43-51.
- [10] HAGEN L, KAHNG A. New spectral methods for ratio cut partitioning and clustering[J]. *IEEE Trans. Comput.-Aided Des.*, 1992, 11(9): 1074-1085.
- [11] SZLAM A, BRESSON X. Total Variation and Cheeger Cut[C]// *Proceeding of the 27th International Conference on Machine Learning*. 2010: 1039-1046.
- [12] SHI J, MALIK J. Normalized cuts and image segmentation[J]. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2000, 22(8): 888-905.
- [13] SZLAM A, BRESSON X. A total variation-based graph clustering algorithm for Cheeger ratio cuts[C]// *International Conference on Machine Learning*. 2010: 1039-1046.
- [14] BRESSON X, LAURENT T, UMINSKY D, et al. Convergence and energy landscape for Cheeger cut clustering[J]. *Advances in Neural Information Processing Systems*, 2012: 25: 1394-1402.
- [15] BRESSON X, LAURENT T, UMINSKY D, et al. An adaptive total variation algorithm for computing the balanced cut of a graph[J]. *arXiv preprint arXiv:1302.2717*, 2013.
- [16] BRESSON X, LAURENT T, UMINSKY D, et al. Multiclass total variation clustering[J]. *Advances in Neural Information Processing Systems*, 2013, 26: 1421-1429.
- [17] BRESSON X, TAI X C, CHAN T F, et al. Multi-class transductive learning based on L1 relaxations of Cheeger cut and Mumford-Shah-Potts model[J]. *Journal of Mathematical Imaging and Vision*, 2014, 49(1): 191-201.
- [18] Hu H, LAURENT T, PORTER M A, et al. A method based on total variation for network modularity optimization using the MBO scheme[J]. *SIAM Journal of Applied Mathematics*, 2013, 73(6): 2224-2246.

(上接第 173 页)

标处在较复杂的环境下时,每帧图像的处理速度平均为 0.27 s,能够有效保证视频的处理速度,因此所提算法在稳定性和鲁棒性方面都取得了较理想的成果。

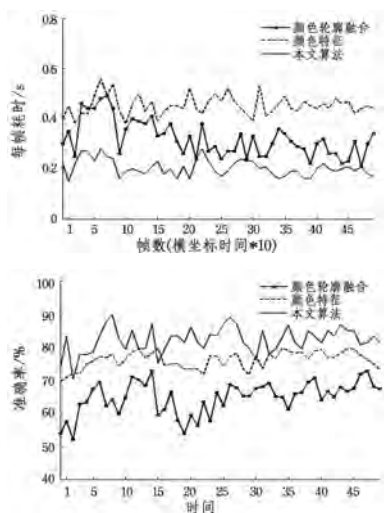


图5 跟踪算法的精确度

结束语 本文提出了一种基于预测的多特征融合的 Mean-Shift 算法,其将颜色与纹理特征进行动态融合,再用二次多项式对目标位置进行预测。在每一帧图像中,运用基于颜色的 Mean-Shift 算法与基于 Gabor-LBP 的 Mean-Shift 算法对目标模型进行计算,同时结合二次多项式预估的目标位置,通过阈值的规范得到最后的目标位置。本文提出的算法一方面能够有效克服颜色复杂、亮度突变导致的跟踪失败,另

一方面通过预测能够有效解决遮挡问题,与一些经典的 Mean-Shift 算法相比,在速度和准确率上都有一定的提升。

参考文献

- [1] 王永忠,梁彦,赵春晖,等. 基于多特征自适应融合的目标跟踪方法[J]. *自动化学报*, 2008(4): 393-399.
- [2] 黄伟国,顾超,尚丽,等. 基于轮廓分层描述的目标识别算法研究[J]. *电子学报*, 2015, 43(5): 854-861.
- [3] 刘明华,汪传生,王宪伦. 基于多特征自适应融合的均值迁移目标跟踪算法[J]. *光电子·激光*, 2015, 26(8): 1583-1592.
- [4] 田莘. 基于 Mean-Shift 算法的目标跟踪问题研究[D]. 西安:西安科技大学, 2010.
- [5] 刘明华. 复杂环境下基于多特征融合的目标跟踪关键技术研究[D]. 青岛:青岛科技大学, 2015.
- [6] 尹宏鹏. 基于计算机视觉的运动目标跟踪算法研究[D]. 重庆:重庆大学, 2009.
- [7] 李鹏飞. 目标跟踪算法的研究与系统构建[D]. 南京:南京师范大学, 2016.
- [8] 袁广林,薛模根,韩裕生,等. 基于自适应多特征融合的 Mean shift 目标跟踪[J]. *计算机研究与发展*, 2010, 47(9): 1663-1671.
- [9] 危自福. 视频目标检测与跟踪关键技术研究[D]. 西安:空军工程大学, 2010.
- [10] CHENG Y. Mean-shift, Mode Seeking, and Clustering [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1995, 17(8): 790-799.
- [11] ZHAO G, PIETIKAINEN M. Local binary pattern descriptors for dynamic texture recognition[J]. *Pattern Recognition*, 2016, 2: 211-214.