Half Quadratic Analysis for Mean Shift: with Extension to A Sequential Data Mode-Seeking Method

Xiaotong Yuan, Stan Z. Li
Center for Biometrics and Security Research & National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Science
Beijing, China, 100080
{xtyuan, szli}@nlpr.ia.ac.cn

Abstract

Theoretical understanding and extension of mean shift procedure has received much attention recently [8, 18, 3]. In this paper, we present a theoretical exploration and an algorithm development on mean shift. In the theory part, we point out that convex profile based mean shift can be justified from the viewpoint of half-quadratic (HQ) optimization. Such analysis facilitates the convergence study and uni-mode bandwidth selection for the latest variation, annealed mean shift [18]. In the algorithm development part of this paper, we extend annealed mean shift inside our HQ framework to a novel method, namely adaptive mean shift *(*Ada-MS*), to detect multiple data modes sequentially from an arbitrary starting point in linear running time. To validate the performance, we couple the investigation with two applications: image segmentation and color constancy. Extensive experiments show that the proposed method is time efficient and initialization invariant.*

1. Introduction

Mean shift is a density mode-seeking technique [9, 4, 5]. As a nonparametric iterative procedure, mean shift algorithms use kernels to compute the weighted average of the observations within a smoothing window. This computation is repeated until convergence is attained at a local density mode. This way, the density modes can be elegantly located without explicitly estimating the density. Among different kernels, the special case of profile based kernels are mostly studied [5, 8]. Mean shift algorithm is widely used in computer vision applications, including tracking [7] and image segmentation [20]

Despite the popularity of mean shift, few attempts have been made since Cheng [4] to understand the procedure theoretically. Cheng [4] shows that mean shift is fundamentally a gradient ascent algorithm with an adaptive step size. Fashing *et al*. show the connection between mean shift and the Newton optimization algorithm [8]. They also find that mean shift is actually a quadratic bound optimization both for stationary and evolving sample sets. Carreira-Perpinan [3] proves that mean shift is equivalent to EM algorithm when kernel is Gaussian, and the quotient-convergence rate is generally linear. One inherent drawback for mean shift is that it can only be used to find local modes. Shen *et al*. developed a multi-bandwidth procedure, namely *annealed mean shift*, to solve global density mode localization problem [18]. Since kernel density estimation is actually a M-estimator [5], the annealed mean shift can be viewed as a special case of deterministic annealing based robust M-estimation [12].

In this paper, we present a theoretical exploration and an algorithm development on mean shift. The motivation for the theoretical part is to justify the current understanding of convex profile based mean shift from the half-quadratic (HQ) optimization viewpoint. HQ is a standard optimization technique in convex analysis [17, 1]. By introducing *dual variables*, non-quadratic convex objective functions can be maximized in a quadratic-like way inside the HQ framework. We show, in the case of convex profiles, that mean shift is actually HQ optimization for density mode detection. This implies that mean shift algorithm converges and the rate is at least linear. We also explicitly give an analytic form of upper-bound for uni-mode promising kernel bandwidth, which makes the bandwidth initialization for annealed mean shift [18] more accurate and operable.

In the algorithm development part, we extend annealed mean shift to a novel sequential method, namely *Ada-MS*, for multiple data modes seeking from arbitrary starting point. This is achieved by making full use of the dual variables introduced by HQ analysis to re-weight all the samples adaptively, guiding the search for remaining significant modes. The time complexity for multiple data modes seeking is accelerated from quadric to linear, and

978-1-4244-1631-8/07/$25.00 ©2007 IEEE

this is compared to commonly used exhaustive initialization method [5].

We apply the results of the Ada-MS work to two applications: image segmentation and color constancy, validated using real-world data. For image segmentation, a comparative test shows that Ada-MS is at least twice as faster as IFGT [20], which to our knowledge is one of the fastest image segmentation method based on mean shift. We then point out that, although derived from the kernel density mode-seeking problem, Ada-MS is also applicable to some other kernel based optimization problems. Convex kernel based linear color constancy is adopted as an example, and comparison to the state of the art [13] is done to clearly demonstrate advantages of Ada-MS over EM solution. At the same time, by piecewise linear assumption, the nonlinear color constancy problem can also be approximately solved in the Ada-MS framework.

The remainder of the paper is organized as follows. In section 2 we briefly survey the background of mean shift procedure. In section 3, we derive convex profile based mean shift procedure with HQ optimization formulation and discuss its convergence property. An up-bound for unimode promising bandwidth is then elegantly derived for annealed mean shift. In section 4, we develop Ada-MS algorithm for fast multiple data modes seeking and test its numerical performance. In section 5, we evaluate practical performance of Ada-MS by two groups of experiments. We conclude our work in section 6.

2. Background of Mean Shift Algorithm

One of the most popular nonparametric density estimators is kernel density estimation. Given a Data set X with N data points $x_i, i = 1, ..., N$, drawn from a population width density function $f(x), x \in \mathbb{R}^D$, the general multi-variable kernel density estimate at x is defined by

$$\hat{f}_K(x) = \frac{1}{N}\sum_{i=1}^{N} w_i K_{\mathbf{H}}(x - x_i) \tag{1}$$

where $K_{\mathbf{H}}(x) = |\mathbf{H}|^{-\frac{1}{2}} K(\mathbf{H}^{-\frac{1}{2}}x)$. Here, $K(\cdot)$ is a kernel function with a symmetric positive defined bandwidth matrix $\mathbf{H} \in \mathbb{R}^{D\times D}$. Sample prior weight w_i satisfies $\sum_{i=1}^{N} w_i = 1$. Employing the profile definition, the general kernel density estimator becomes

$$\hat{f}_K(x) = \frac{c_k}{N|\mathbf{H}|^{\frac{1}{2}}}\sum_{i=1}^{N} w_i k(M^2(x, x_i, \mathbf{H})) \tag{2}$$

where $k(\cdot)$ is a profile of the kernel $K(\cdot)$, $M^2(x, x_i, \mathbf{H}) = (x - x_i)^T\mathbf{H}^{-1}(x - x_i)$ is the Mahalanobis distance from x to x_i, w_i is the weight for sample x_i, and c_k is a normalization constant. The mean shift optimization procedure is performed by setting gradient equal to zero and the incremental iteration scheme is obtained immediately:

$$x \leftarrow \frac{\sum_{i=1}^{N} w_i g(M^2(x, x_i, \mathbf{H}))x_i}{\sum_{i=1}^{N} w_i g(M^2(x, x_i, \mathbf{H}))} \tag{3}$$

where $g(x) = -k^{'}(x)$ and $k(\cdot)$ is called as the shadow of the profile $g(\cdot)$ [5].

3. Half Quadratic Analysis For Mean Shift

3.1. Basic Descriptions

In this section, we will build on the convex profile based mean shift as HQ optimization. The results follow directly from standard material in convex analysis (e.g. [17, 14]) and we will omit the technical proofs for page limit. All the conditions we impose on profile $k(\cdot)$ are summarized as below:

1. $k(x)$ is a continuous monotonously decreasing and strictly convex function
2. $\lim_{x\to 0^+} k(x) = \beta > 0$, $\lim_{x\to+\infty} k(x) = 0$
3. $\lim_{x\to 0^+} k^{'}(x) = -\gamma < 0$, $\lim_{x\to+\infty} k^{'}(x) = 0$, $\lim_{x\to+\infty}(-xk^{'}(x)) = \alpha < \beta$
4. $k'(x)$ is continuous with finite discontinuous points.

The following theorem 3.1 founds the base for optimizing density function in a half quadratic way.

Theorem 3.1 *Let $k(.)$ be a profile satisfying all above conditions, then there exists a strictly monotonously increasing concave function $\varphi : (0, \gamma) \mapsto (\alpha, \beta)$, such that*

$$k(M^2(x, x_i, \boldsymbol{H})) = \sup_p(-pM^2(x, x_i, \boldsymbol{H}) + \varphi(p))$$

and for a fixed x, the supmum is reached at $p = -k^{'}(M^2(x, x_i, \boldsymbol{H}))$.

To study kernel density estimator (2), we introduce a new objective function $F : \mathbb{R}^D \times (0, \gamma)^N \mapsto (0, +\infty)$

$$\hat{F}(x, \mathbf{p}) = \frac{c_k}{N|\mathbf{H}|^{\frac{1}{2}}}\sum_{i=1}^{N} w_i(-p_i M^2(x, x_i, \mathbf{H}) + \varphi(p_i)) \tag{4}$$

where $\mathbf{p} = (p_1, ..., p_N)$. According to theorem 3.1, we get

$$\hat{f}_K(x) = \sup_{\mathbf{p}}(\hat{F}(x, \mathbf{p})) \tag{5}$$

It is straight forward to see that

$$\max_x \hat{f}_K(x) = \max_x \sup_{\mathbf{p}}(\hat{F}(x, \mathbf{p})) \tag{6}$$

From (6) we tell that maximizing $\hat{f}_K(x)$ is equivalent to maximizing $\hat{F}(x, \mathbf{p}))$, which is quadratic w.r.t. x when $\mathbf{p}$ is fixed. The maximizer $(\hat{x}, \hat{\mathbf{p}})$ of $\hat{F}$ is calculated by *alternate maximization* as follows (superscript l denotes the time stamp):

$$p_i^l = -k^{'}(M^2(x^{l-1}, x_i, \mathbf{H})), i = 1, ..., N \tag{7}$$

$$x^l = \frac{\sum_{i=1}^N w_i p_i^l x_i}{\sum_{i=1}^N w_i p_i^l} \tag{8}$$

It is obvious to see that the above two-step iterative procedure is equivalent to the mean shift procedure. The *dual variables* $\mathbf{p}$ play the role of $g(\cdot)$ in (3).

3.2. Relation to Bound Optimization and EM

There is an interesting relation with the result of Fashing and Tomasi [8],which builds on the mean shift as a quadratic bound maximization. For a fixed point $x^{l-1} \in \mathbb{R}^D$, denote the quadratic function $\hat{\rho}(x) = \hat{F}(x, \mathbf{p}^l)$ ($\mathbf{p}^l$ is obtained by (7)). We can get from theorem 3.1 and (5) that $\hat{f}_K(x^{l-1}) = \hat{\rho}(x^{l-1})$ and $\hat{f}_K(x) \geq \hat{\rho}(x)$ for $\forall x$, i.e. $\hat{\rho}(x)$ defines a lower bounding function for $\hat{f}_K(x)$ at point x^{l-1}. It is easy to see that iteration step (8) maximizes quadratic function $\hat{\rho}(x)$. In this way, the proposed HQ analysis can also be viewed as a *quadratic bound optimization* for mean shift. This is tightly related to [8], and the quadratic lower bound given by (15) in [8] can be proved equivalent to $\hat{\rho}(x)$ defined here. The improvement lies in that, instead of locally computing coefficients ((18) in [8]), our HQ based approach allows a "global" bound expression (6), which is suitable for numerical study on convergence and global mode-seeking.

When kernel is Gaussian, our HQ optimization is also equivalent to the recent result of Carreira-Perpinan [3], which shows that Gaussian mean shift is an EM algorithm on a properly defined Gaussian-mixture density model, when trying to fit a sample at the origin. In our notation, in the E-step, the posterior distribution $p(i|x^{l-1}) \propto p_i^l$ and the expectation w.r.t. the current posterior is proportional to the above defined $\hat{\rho}(x)$. In the M-step, the new mode x^l is obtained from old x^{l-1} by (8).

3.3. Convergence Study

The HQ formulation also facilities the convergence proof of mean shift procedure. We first introduce two concepts that will be used in the context.

Definition The *Mahalonobis diameter* of data set X is defined as $D_{\mathbf{H}}(X) = \max\{\sqrt{M^2(x_i, x_j, \mathbf{H}))}|x_i, x_j \in X\}$.

Definition The *convex-hull* of data set X is defined as $S(X) = \{\sum_{i=1}^N s_i x_i | \sum_{i=1}^N s_i = 1, s_i > 0\}$.

Proposition 3.1 *Denote $\hat{F}^l = \hat{F}(x^{l-1}, \boldsymbol{p}^l)$, the sequences $\{\hat{F}^l, l = 1, 2, ...\}$, $\{x^l, l = 1, 2, ...\}$ and $\{\boldsymbol{p}^l, l = 1, 2, ...\}$ generated by (7) and (8) converge on $S(X)$.*

The proof is trivial, we omit it from this conference paper for page limit.

Until now, the convergence rate of mean shift has been scarcely addressed. In [3], it is established that the quotient-convergence order of the Gaussian kernel mean shift is generally linear [1]. On the other hand, the convergence rate of HQ is deeply studied in literature [1, 14], rather relies on root-convergence [2] factors. Since our work bridges the gap between mean shift and HQ optimization, the discussion on convergence rate for mean shift is facilitated. Denote spectral $\sigma = \lambda_{max}\left(\mathbf{I} + \frac{\mathbf{H}\bigtriangledown^2 \hat{f}_K(\hat{x})}{2\sum_{i=1}^N \hat{p}_i}\right)$, $\lambda_{max}(.)$ is the largest eigenvalue of a given square matrix. $\mathbf{I}$ is the identity matrix. If kernel band matrix $\mathbf{H}$ is isotropic ($\mathbf{H} = h^2\mathbf{I}$), then $\sigma < 1$ since $\hat{f}_K$ is negatively defined at local maximizer $\hat{x}$. The main result on convergence rate for mean shift is:

Proposition 3.2 *The root-convergence of convex kernel based mean shift scheme (7) and (8) is at least linear with rate σ*

The proof is based on linear convergence theorem [15].

3.4. Global Mode-Seeking

Since the standard mean shift is essentially a gradient ascending method, it will converge to local maximum. Recently, Shen et al. [18] developed a multi-bandwidth variation of mean shift, namely *annealed mean shift*, to solve global kernel density mode-seeking problem. In this work, we revisit annealed mean shift inside the HQ optimization framework (see algorithm 1 for a formal description in our notation). The key contribution is that we explicitly give an up bound of the critical uni-mode-promising bandwidth (proposition 3.3), which makes the bandwidth initialization more accurate and operable in practice.

Here, we just equivalently discuss uni-mode property of (4). We consider a special case that $\mathbf{H} = \eta^2\mathbf{H}_0$, $\mathbf{H}_0$ is a fixed matrix, e.g. the second moment of data set X. Then (4) can be rewritten as

$$\hat{F_\eta}(x, \mathbf{p}) = \frac{c_k}{N|\mathbf{H}_0|^{\frac{1}{2}}\eta^{D+2}} \sum_{i=1}^N w_i(-p_i M^2(x, x_i, \mathbf{H}_0) + \eta^2\varphi(p_i)) \tag{9}$$

Proposition 3.3 shows that (9) is concave (hence is unimode) on the convex-hull $S(X)$ for a large enough η.

[1] We say that the convergence in quotient is linear if there exists a $\varepsilon \in (0, 1)$ such that $\|x^{l+1} - \hat{x}\| \leq \varepsilon\|x^l - \hat{x}\|$ for all l sufficiently large.

[2] We say that the convergence in root is linear if there exists a $\sigma \in (0, 1)$ such that $\sup_{l\to\infty} \|x^l - \hat{x}\|^{1/l} \leq \sigma$.

Proposition 3.3 *One sufficient condition guarantees that* $\hat{F}_\eta(x,\boldsymbol{p})$ *is concave on* $S(X)$ *is*

$$\eta > \left(2\sup_v\left(-\frac{k''(v)}{k'(v)}\right)\right)^{\frac{1}{2}} D_{\boldsymbol{H}_0}(X) \tag{10}$$

Remark Given the η setting condition presented in proposition 3.3, it is easy to see from (6) that the estimated density function $\hat{f}_K(x)$ is also uni-mode. Further more, proposition 3.3 implies that the *critical uni-mode-promising bandwidth*, defined as η_{crit}=inf$\{\eta_{um} > 0$: $\hat{f}_K$ is uni-mode for all $\eta > \eta_{um}\}$ [16], is up bounded by the right side of inequality (10). We give some commonly used kernels and their corresponding uni-mode-promising bandwidths in table 1 to further clarify proposition 3.3.

Table 1. Kernels and bandwidths

$k(x)$	uni-mode-promising bandwidth
$e^{-x/2}$	$> D_{\mathbf{H}_0}(X)$
$\frac{1}{1+x}$	$> 2D_{\mathbf{H}_0}(X)$
$\frac{\pi}{2} - \arctan(x)$	$> \sqrt{2}D_{\mathbf{H}_0}(X)$

From proposition 3.3 and 3.1, we can tell that if η is large enough , then from any initial estimation, the two-step iteration (7) and (8) will converge to a unique maximizer of the over-smoothed density function. We may then gradually decrease η (in this paper, by multiplying a constant $\theta \in (0,1)$) and run the same iterations again, taking the previous maximizers as current initializations. This procedure is repeated until a certain termination condition is met (e.g. η_m reaches AMISE optimal bandwidth [6] for the considered data set), and the final obtained maximizer is very likely to be the global maximum point of density function. In this way we revisit the annealed mean shift stated in [18].

Algorithm 1 Annealed Mean shift

1: $m \leftarrow 0$, Initialize η_m satisfying the condition presented in proposition 3.3
2: Randomly select an initial starting location from $S(X)$
3: **while** Terminate condition is not met **do**
4: Run the iteration (7) and (8) till converge.
5: $m \leftarrow m + 1$
6: $\eta_m \leftarrow (\eta_{m-1} * \theta)$.
7: Initialize x and **p** with the maximizers obtained in 4.
8: **end while**

To summarize the theoretical exploration so far, we justify the understanding of convex profile based mean shift on the viewpoint of half-quadratic optimization. The state of the art variation, annealed mean shift, is more rigourously rivisited in our HQ framework. In the following subsections, we will further develop a novel multiple data modes seeking method based on above analysis. For presentation clarity, we denote x^* and $\mathbf{p}^*$ be the convergent points reached in algorithm1, and η^* be the corresponding bandwidth. We also call the global maximizer x^* to be *global data mode* (*GDM*) of set X associated with current prior weights **w**.

4. Extension: Ada-MS For Sequential Data Mode-Seeking

In many applications, e.g. discontinuity preserving smoothing and image segmentation [5], mean shift procedure is often used for multiple data modes seeking and clustering. Typically, to detect all the significant modes, the basic mean shift algorithm stated in Section 3.1 should be run multiple times with initializations that cover the entire feature space [5]. Given N data sample points, direct estimation of local modes from all these initializations will take $\mathcal{O}(N^2)$ evaluation. For large data set, such an exhaustive mechanism leads to severe requirements for computational time and/or storage. Yang *et al*. [20] accelerated the mode-seeking speed to linear running time by using improved fast gaussian transform (IFGT). Although very efficient for the special case of Gaussian kernel, IFGT seems difficult to be generalized for other convex kernels. At the same time, it remains an exhaustive initialization scheme. In this section, for general convex kernels, we develop a novel method, namely *adaptive mean shift* (*Ada-MS*), to sequentially detect the significant data modes in linear time complexity. The method can be viewed as an extension of annealed mean shift inside our HQ analysis framework.

4.1. Algorithm Description

The core idea of Ada-MS algorithm is to find multiple data modes one after another by adaptively changing the prior weight vector **w**. Points closer to currently found modes receive lower weights, allowing the others to guide the search in the next iteration. With current sample prior weight, we run annealed mean shift to locate *GDM* x^*, and then taking it as starting point to find local maximizer $x^{*'}$ for the density function (2) estimated under equal prior weight (remind that our purpose is to find the modes for original data set). Dual variable vector **p** is calculated as $p_i = -k'(M^2(x^{*'}, x_i, \eta^{*2}\mathbf{H}_0)$, $i = 1, ..., N$. We then reweight all the samples by $w_i \leftarrow w_i/(1 + p_i)$ and normalize them. This procedure is repeated until some ever-found mode reappears. Suppose L modes are eventually obtained, taking them as centers, the clustering can be done by naive nearest-neighbor scheme. The running time is accelerated to linear complexity $\mathcal{O}(LN)$ ($L \ll N$). The formal and detailed description of Ada-MS is given in algorithm 2.

Algorithm 2 Ada-MS for Mode-Seeking and Clustering

Initialization: Start with weights $w_i^0 = 1/N$, $i = 1, ..., N$ and $l = 0$. Set mode set $\mathscr{L} = \varnothing$.
while 1 **do**
  ***GDM* Estimation:** Find *GDM* x^* by annealed mean shift for $\hat{f}_K(x)$ estimated under prior weight $\mathbf{w}^l$.
  Mode Locolization: Starting from x^*, find the local maximizer $x^{*'}$ by mean shift for $\hat{f}_K(x)$ estimated under η^* and $\mathbf{w}^0$.
  if $x^{*'} \in \mathscr{L}$ **then**
    break
  else
    Dual Variables: Get $p_i = -k'(M^2(x^{*'}, x_i, \eta^{*2}\mathbf{H}_0))$.
    Sample Reweight: Set $w_i^{l+1} \leftarrow w_i^l/(1+p_i)$. Normalize $w_i^{l+1} \leftarrow w_i^{l+1}/\sum_i w_i^{l+1}$
    $l \leftarrow l + 1$
  end if
end while
Clustering: the data set X are grouped via naive nearest-neighbor algorithm, taking the L $(= \|\mathscr{L}\|)$ modes in $\mathscr{L}$ as centers

4.2. Numerical Test

We give in this section an 1D data mode-seeking experiment to more clearly illustarte the numerical procedure of the Ada-MS. Galaxies data set (1-D, size 11264) from [10] is adopted for this test. Ada-MS successfully found the 3 significant modes appeared in this data set.

***GDM* Estimation:** Firstly, we estimate the *GDM* under initial equal prior weights. The kernel profile used here is $k(x) = e^{-x/2}$. Fig. 1(a) shows that *GDM* is successfully located with a rough 4-step iteration.

Sequential mode-seeking: fig.1(b) shows the multiple modes seeking results by Ada-MS. Estimated density curves under sequentially changed sample prior weights are shown. Corresponding *GDM*'s and data modes are gradually located in this procedure. Eventually, three significant modes are all correctly located. The sample prior weight curves (from initial one to the 4th re-weighting iteration) are shown in fig.1(c) $\sim$1(f)

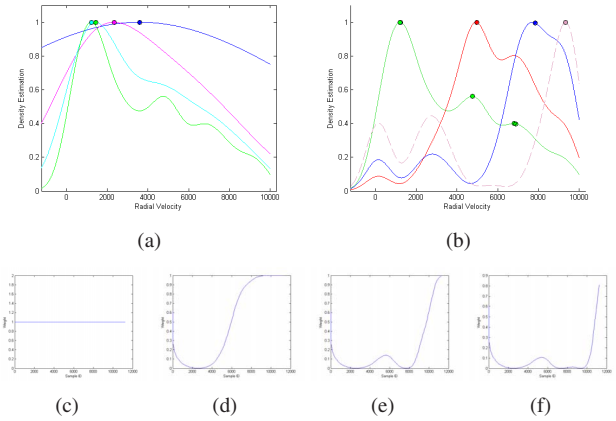

Figure 1. (a) Annealed mean shift for *GDM* detection. In this case, the estimated initial bandwidth by proposition 3.3 is set to be $\eta_0 = 8.3470$ and shrinking factor is $\theta = 0.5$. Curves from outside to inside indicate the evolving process with successively decreasing bandwidths (η_0, η_1, η_2, η_3). The evolution of the modes is clearly shown and the *GDM* is located without being distracted by local modes. (b) Multiple modes seeking result by Ada-MS. The green curve is the estimated density curve under initial equal sample weights, while the red, blue and pink curves are those under sequentially re-weighted prior. The *GDM*'s are marked as circles on the corresponding curves. The estimated local data modes are shown on the initial green density curve. On the 4th sample re-weighting iteration (pink dotted curve), the estimated mode is overlapped with the 3rd mode, hence the iteration stopped with eventually three modes found. For each iteration, the initial bandwidth η_0 estimated by proposition 3.3 is 8.3470. The corresponding sample prior weight curves are shown in (c) $\sim$ (f).

5. Applications of Ada-MS

To further evaluate the practical performance of the proposed Ada-MS data mode-seeking method, we couple this investigation with two real-world applications: image segmentation and color constancy.

5.1. Static Image Segmentation

Firstly, we present an experiment for static image segmentation, aiming to show the computational efficiency of the proposed Ada-MS on high dimensional data set. Image segmentation is a fundamental component in many computer vision applications, and can be addressed as an image data clustering problem [11]. We use several test images from the Berkeley Segmentation Dataset [2] and from [20] for evaluation. We adopt the L*U*V* color features to form a 3-dimensional raw feature space. Ada-MS with Gaussian profile is applied to all these test images in such a 3D color space. The code is written in C++ with Matlab interface, and run on a 3.0G Hz P4 CPU. We also assume that $\mathbf{H}_0$ is isotropic. The results are shown in fig.3. The running time of the algorithm 2 in seconds and the sizes of the images are shown in table 2. As a comparison, we also list in table 2 the corresponding computational time by the above mentioned linear time algorithm IFGT [20]. On our test environment, IFGT is already twice faster than its original version developed in [20]. From the comparison, we can see that our Ada-MS based method is more efficient than IFGT. This is mainly due to the fact that IFGT remains an exhaustive searching mechanism while our method is a sequential one.

5.2. Color Constancy

The development and applications of Ada-MS so far are just restricted in the context of kernel density mode-seeking

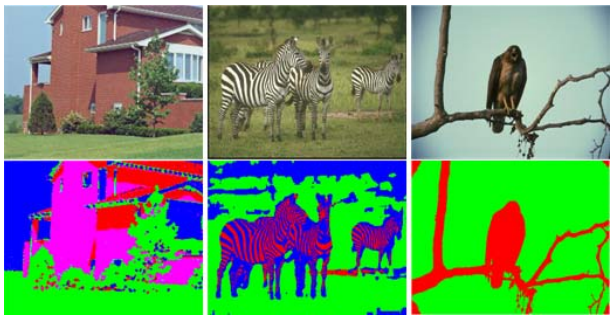

Figure 2. Image segmentation results: Column $1 \sim 3$: *House* image ($\eta_0 = 3.292$, $\theta = 0.6$, $L = 4$), *Zebra* image ($\eta_0 = 5.269$, $\theta = 0.6$, $L = 3$), and *Bird* image ($\eta_0 = 4.341$, $\theta = 0.6$, $L = 2$).

Table 2. Image sizes v.s. the running time

		House	*Zebra*	*bird*
size		255×192	481×321	481×321
Time (s)	IFGT [20]	2.063	6.516	6.640
	Ada-MS	1.062	3.406	1.485

problem. In this application, we relieve this restriction and show how it can be used for other kernel based optimization problems. We take color constancy problem as an example and use Ada-MS to optimize the diagonal render model discussed in [13]. The main purpose of this group of experiments is to validate that Ada-MS is an initialization invariant optimization framework, thus is superior to EM solution adopted in [13].

As is well known that light sources, shadows, transducer non-linearities, and camera processing can all affect the final image of a scene. Color constancy, which addresses the variability of images due to above photic parameters, is an important problem in machine vision. A large body of work (see [19] and the refs there in) has been presented for this research topic. Recently, [13] presented a diagonal rendering model for outdoor color compensation and classification problem, in which only one image containing the color samples under a certain "canonical" illumination is needed for training. The trained colors seen under different illuminations in the test image can be robustly recognized via MAP estimation. The key assumptions for this model are:

- One hand-labeled image is available for training the class-conditional color distributions under the "canonical" illuminant.
- The class-conditional color surface likelihood under the canonical illumination is a Gaussian density, with mean μ_j and covariance Σ_j
- The illuminant-induced color transformation in the test image can be modeled as $F(C_i) = C_i\mathbf{d}$, where $\mathbf{d} = (d_1, d_2, d_3)^T$ is the color render vector to be determined. $C_i = diag(r_i, g_i, b_i)$ is a diagonal matrix that stores the observed RGB colors for pixel i in the test image.

Suppose S color surfaces with distributions $y_j \sim \mathcal{N}(\mu_j, \Sigma_j)$,$j = 1, ..., S$ are trained. Also, assume given a test image with N pixels C_i, $i = 1, ..., N$, which contains L illuminants linearly parameterized by vectors $\mathbf{d}_l$, $l = 1, ..., L$. The goal is to estimate these $\mathbf{d}_l$ from image data (both training and test) and then get the assignments of surface class labels $j(i)$ and illuminant type labels $l(i)$ for each pixel i according to:

$$(j(i), l(i)) = \arg\min_{j,l}(dist(C_i\mathbf{d}_l, y_j)) \qquad (11)$$

$dist(\cdot)$ is some properly selected distance measurement metric (e.g. Mahalanobis distance in this work).

Due to the advantage of fewer training images requirements, we also adopt this render model for our color constancy application. The main difference between our solution and that of [13] lies in the definition of objective function and the associated optimization method to estimate parameters $\mathbf{d}$. In [13], compensation error based image likelihood and model priors are integrated into a MAP formulation and possible existing $\mathbf{d}$'s are optimized through EM algorithm. This algorithm works well when L is known a prior and all the render vectors are properly initialized. However, in practice, such information is not always available or accurate. We successfully overcome this drawback by properly defining the following convex kernel based criterion function to measure the illumination compensation accuracy:

$$\hat{f}_K(\mathbf{d}) = \sum_{i=1}^{N}\sum_{j=1}^{S} w_{ij}k(M^2(C_i\mathbf{d}, \mu_j, \mathbf{H}_j)) \qquad (12)$$

where $\mathbf{H}_j = \eta^2\Sigma_j$ and w_{ij} is the prior weight for pixel i belonging to color surface j. Profile $k(.)$ also satisfies all the conditions provided in 3.1. The larger (12) is, the better test image is compensated by $\mathbf{d}$.

Ada-MS algorithm is applied to find the significant modes of criterion function (12). A slightly revision is that the two-step iteration (7) and (8) is now derived as

$$p_{ij}^l = -k'(M^2(C_i\mathbf{d}^{l-1}, \mu_j, \mathbf{H}_j))$$

$$\mathbf{d}^l = \left[\sum_{i=1}^{N}\sum_{j=1}^{S} w_{ij}p_{ij}^l C_i^T\mathbf{H}_j^{-1}C_i\right]^{-1}\left[\sum_{i=1}^{N}\sum_{j=1}^{S} w_{ij}p_{ij}^l C_i^T\mathbf{H}_j^{-1}\mu_j\right]$$

Similar results to proposition 3.3 on uni-mode-promising bandwidth can also be derived .

Suppose eventually L number of modes $\mathbf{d}_l$ are found via Ada-MS, we may naturally view these modes as significant illumination transformations in the scene. Color and illuminant type classification can be done according to (11).

We present two sets of experiments on color compensation and classification for real scene images to show the performance of our method. Gaussian kernels are used in these experiments.

The first experiment is done to show the initialization invariant property of our algorithm. For comparison purpose, we adopt one set of image data used in [13]. Fig.3(a) is the training image with selected sample colors under "canonical" light. The test image is shown in fig.3(b). Fig.3(c) and 3(d) are compensation results achieved by [13] with EM based optimization, from starting point P_1 and P_2 (see table3) separately. It is obviously to see that fig.3(c) is much more satisfying than fig.3(d), hence the algorithm is highly initialization relevant. The results by our Ada-MS mode-seeking algorithm are shown in fig.3(e) ∼3(h). Detailed numerical comparison results are listed in table 3. Note that, for Ada-MS, the calculated $\mathbf{d}_3 = (2.054, 1.565, 1.401)$ is equivalent to $\mathbf{d}_2$, hence we get $L = 2$, which coincides the observation that two significant illuminant (light and shadow) exist in the scene. Fig. 4 more clearly visualize the global mode-seeking and sample re-weighting process during optimization.

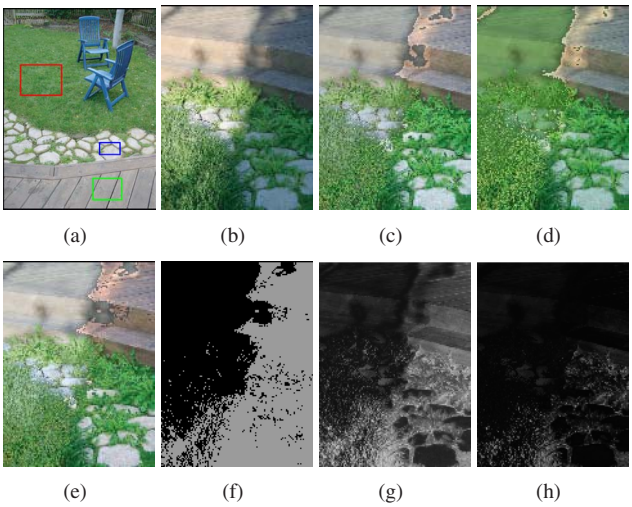

Figure 3. A comparison example with EM based method [13]. (a) training image; (b) test image; (c) and (d): compensation results by EM optimization [13], from starting point P_1 and P_2 separately; (e) and (f): the compensation and illumination classification results by our Ada-MS ($\eta_0 = 2.478$ and $\theta = 0.6$, $L = 2$), which are invariant to starting point; (g) and (h): weight maps (for each pixel $w_i = \sum_{j=1}^{S} w_{ij}$) for the 2nd and 3rd sample re-weighting iteration.

The second experiment will show the ability of our method to handle non-linear illumination changes based on current linear render model. To do this, we make piecewise linear assumption to approximate the general nonlinear cases. Our method can automatically find the transformation vectors for each linear piece. We give here one experiment on a pair of "map" images to validate this interesting property. We used Canon A550 with automatic exposure, taking care to compensate for the camera's gamma setting. The training image fig.5(a) and test image fig.5(b) are shot under two very different camera settings. The selected 6 sample colors from the training image and their ground truth counterparts in the test image are shown in fig.5(c) (left). To test whether the illumination change in the test image is linear or not, we calculate the ground truth transformation vectors for the sample colors and plot them in fig.5(c)(right). Obviously two clusters (bounded by dotted ellipses) appear from these vectors, thus the illumination change is nonlinear. A reasonable assumption is that such a change is piecewise linear and we may just feed the image data into Ada-MS to let it find all the significant pieces sequentially, from arbitrary initializations. EM based method [13] can hard to achieve this goal simply because the number of pieces and accurate initialization for each linear piece is required to be known a prior, which is not always available in practice. On the other hand, our Ada-MS successfully found two vectors $\mathbf{d}_1 = (0.638, 0.836, 1.611)$ and $\mathbf{d}_2 = (0.765, 0.977, 3.022)$ with arbitrary starting points (parameters are set as $\eta_0 = 1.934$ and $\theta = 0.5$). The image results are shown in fig.5(d) ∼5(h).

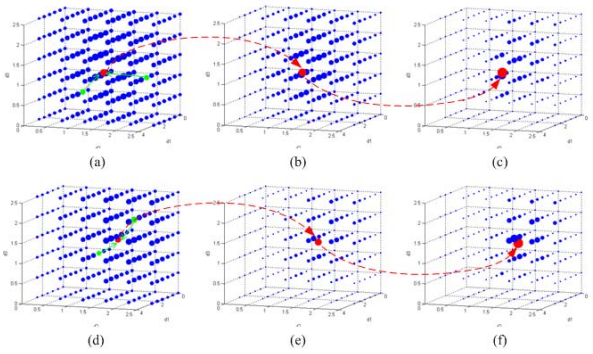

Figure 4. Visualization of the compensation accuracy function (12) in domain $(0, 2.5)^3$ under different prior sample weight. (a) ∼ (c) are for initial equaling prior weights while (d) ∼ (f) are for the 2nd round of re-weighted prior. The radius of each dot indicates compensation accuracy at that point. Graphes from left to right indicate the evolving process with successively decreasing bandwidths (η_0, η_1, η_4). Global modes (in red) are located via Ada-MS from given initializations (in green).

Table 3. Numerical results, Ada-Mean-Shift vs. EM

	$\mathbf{d}_1$	$\mathbf{d}_2$
Starting point $P1$	(1.0,1.0,1.0)	(2.0,2.0,2.0)
Result by EM [13]	(0.946,0.988,1.072)	(2.202, 1.737,1.546)
Result by Ada-MS	**(0.899,0.968,1.038)**	**(2.057,1.567,1.404)**
Starting point $P2$	(0.5,0.5,0.5)	(1.0,2.0,1.0)
Result by EM [13]	(0.499,0.754,0.508)	(1.627, 1.486,1.247)
Result by Ada-MS	**(0.899,0.968,1.038)**	**(2.057,1.567,1.404)**

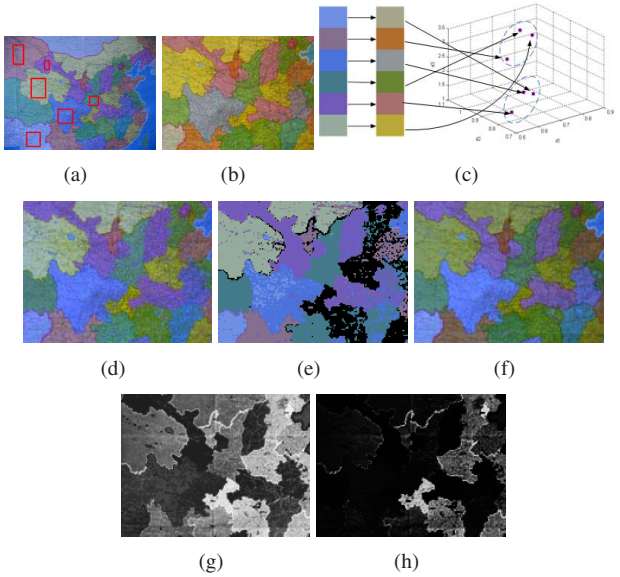

Figure 5. Piecewise linear color constancy. (a)training image; (b)test image; (c) left: 6 selected sample colors and their ground truth counterparts in the test image; right: the ground truth transformation vectors for the 6 sample colors; (d) and (e): color compensation and color classification results. The black part in (e) represents unseen colors in the test image. (f): color compensation result by render vector $\mathbf{d}_1$ only, which obviously introduces very large compensation error, visually. (g) and (h): weight maps for the 2nd and 3rd sample re-weighting iteration.

6. Conclusion

We investigated into the mechanism behind the mean shift procedure and explained convex profile based mean shift in terms of half-quadratic (HQ) optimization and discussed its convergence rate. We applied the HQ analysis to solve the uni-mode bandwidth selection problem associated with annealed mean shift [18]. We further developed Ada-MS method for fast multiple data modes seeking, from arbitrary starting point. The computational complexity is reduced from quadratic to linear. Extensive experiments validate the time efficiency and initialization invariance property of Ada-MS.

7. Acknowledgement

This work was supported by the following funding resources: National Natural Science Foundation Project #60518002, National Science and Technology Supporting Platform Project #2006BAK08B06, National 863 Program Projects #2006AA01Z192 and #2006AA01Z193, and Chinese Academy of Sciences "100 people project".

References

[1] M. Allain, J. Idier, and Y. Goussard. On global and local convergence of half-quadratic algorithms. *IEEE Transitions on Image Processing*, 15(5):1130–1142, 2006.

[2] Berkeley. *http://www.eecs.berkeley.edu/Research /Projects/CS/vision/bsds/BSDS300/html/dataset/images.html*.

[3] M. Carreira-Perpinan. Gaussian mean-shift is an em algorithm. *IEEE Transactions on Pattern Analysis and Macnine Intelligence*, 29(5):767–776, 2007.

[4] Y. Cheng. Mean shift, mode seeking, and clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(7):790–799, 1995.

[5] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):603–619, May 2002.

[6] D. Comaniciu and P. Meer. An algorithm for data-driven bandwidth selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(2):281–288, Feb 2003.

[7] D. Comaniciu, V. Ramesh, and P. Meer. Real-time tracking of non-rigid objets using mean shift. In *Computer Vision and Pattern Recongition,IEEE International Conference on*, volume 2, pages 142–149. IEEE, 2000.

[8] M. Fashing and C. Tomasi. Mean shift is a bound optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27:471–474, Mar. 2005.

[9] K. Fukunaga and L. Hostetler. The estimation of the gradient of a density function, with application in pattern recognition. *IEEE Transactions on Information Theory*, 21:32–40, 1975.

[10] G.Palumbo, G.Tanzella-Nitti, and G.Vettolani. *Catalogue of Radial Velociies of Galaxies*. New York: Gordron and Breach, 1983.

[11] A. K. Jain, M. N. Murty, and P. Flynn. Data clustering: A review. *ACM Computing Surveys*, 31(3):264–323, 1999.

[12] S. Z. Li. Robustizing robust m-estimation using deterministic annealing. *Pattern Recognition*, 29(1):159–166, 1996.

[13] R. Manduchi. Learning outdoor color classification. *IEEE Transactions on Pattern Analysis and Macnine Intelligence*, 28(11):1713–1723, 2006.

[14] M. Nikolova and M. NG. Analysis of half-quadratic minimization methods for signal and image recovery. *SIAM Journal of Scientific Computing*, 27(3):937–966, 2005.

[15] J. Ortega and W. Rheinboldt. *Iterative Solution of Nonlinear Equations in Several Variables.* New York: Academic, 1970.

[16] P.Hall, M.C.Minnotte, and C.Zhang. Bump hunting with non-gaussian kernels. *The Anneal of Statistics*, 32(5):2124–2141, 2004.

[17] R. Rockfellar. *Convex Analysis*. Princeton Press, 1970.

[18] C. Shen, M. Brooks, and A. van den Hengel. Fast global kernel density mode seeking with application to localization and tracking. In *IEEE International Conference on Computer Vision*, volume 2, pages 1516–1523. IEEE, 2005.

[19] K. Tieu and E. Miller. Unsupervised color constancy. In *NIPS*, 2002.

[20] C. Yang, N. Duraiswami, R.and Gumerov, and L. Davis. Improved fast gauss transform and efficient kernel density estimation. In *IEEE International Conference on Computer Vision*, volume 1, pages 664–671. IEEE, 2003.