Asymmetrical Occlusion Handling Using Graph Cut for Multi-View Stereo

Yichen WEI Long QUAN
Department of Computer Science, Hong Kong University of Science and Technology
Clear Water Bay, Kowloon, Hong Kong
{yichenw,quan}@cs.ust.hk

Abstract

Occlusion is usually modelled in two images symmetrically in previous stereo algorithms which cannot work for multi-view stereo efficiently. In this paper, we present a novel formulation that handles occlusion using only one depth map in an asymmetrical way. Consequently, multi-view information is efficiently accumulated to achieve high accuracy. The resulting energy function is complex and approximate graph cut based solutions are proposed. Our approach complements the theory and extends the applicability of using graph cut in stereo. The experiments demonstrate that the approach is comparable with the state of the art and potentially more efficient for multi-view stereo.

keywords : stereo, graph cut, occlusion

1. Introduction

Stereo vision addresses the problem of obtaining depth information from multiple images taken about the static scene from different viewpoints. Occlusion handling is a challenging problem in stereo matching since correspondences of pixels in occlusion area are not well defined. Previous stereo algorithms either ignore occlusion area, therefore computing a quasi-dense depth map [12], or use various techniques to handle occlusion explicitly [5].

The occlusion handling in multi-view stereo has been discussed in [8]. Several heuristic techniques are proposed to reduce the sensitivity to occlusion, such as shiftable windows and temporal selection of frames. These techniques are not used in our approach since we are focusing on the problem formulation and energy function minimization. However, they may further improve the results.

Dynamic programming approaches [3, 6] work on two images symmetrically. They solve an 1D path-finding optimization problem which is formulated on the corresponding epipolar lines and computes occlusion explicitly. There are two main problems in such approaches. Ordering constraint is usually assumed but it may be violated in practice. It is hard to ensure intra-scanline consistency in the depth map which is not explicitly formulated in the energy function. These problems are avoided in our approach by solving a more general 2D optimization problem.

There are a lot of stereo algorithms using graph cut. We only discuss those [4, 9, 10] which are most related to our approach, as compared in Table 1. Occlusion is not considered in [4] and it depends on the smoothness constraint to achieve good results. The dependency is reduced in [9] by adding a visibility term to the energy function addressing the uniqueness constraint, but the formulation is restricted on two views. The multiple view approach [10] imposes visibility constraints to arbitrarily two views by exploiting the epipolar geometry between them.

Our formulation extends the energy function in [4] to handle occlusion explicitly. The occlusion computing process is similar as in [10]. However, since only one depth map is computed, the formulation is more complicated. The resulting energy function is smaller, more difficult and cannot be minimized by graph cut in general. Instead, approximate minimization techniques are developed and they are shown to be effective in practice.

While the approach in [10] is obviously more general, the significance of proposed approach is two-fold. Theoretically, the proposed approach shows that it is possible to effectively handle occlusion in one depth map using graph cut. This complements the theory and extends the applicability of using graph cut in stereo, as revealed in Table 1.

Practically, the approach in [10] minimizes a large energy function which is a sum of two-view energy function, called an interaction of the two views. The result depends on how those interactions are defined. The best result is obtained, of course, when every two view interaction is included. However, the rapidly increasing time and space complexity limits the maximal number of images that can be used. In experiments, the proposed approach is compared to the approach in [10] in various cases. The results show that they are comparable in accuracy, and the proposed approach is potentially more efficient for multi-view stereo, allowing usage of more images with larger size.

	#input images (#depth maps computed)	handling visibility (occlusion)	energy function	dependency on E_{smooth}
bvz99[4]	n(1)	no	$E_{data} + E_{smooth}$	high
kz01[9]	2(2)	yes	$E_{data} + E_{smooth} + E_{occ}$	medium
kz02[10]	n(n)	yes	$E_{data} + E_{smooth} + E_{visibility}$	low
proposed method	n(1)	yes	$E_{data} + E_{smooth}$	low

Table 1. Comparison of several graph cut stereo algorithms. The methods in [4, 9] do not use epipolar constraint. Consequently, they compute the pixel correspondences, not necessarily the depth maps.

2. Problem Formulation

Suppose we are given n calibrated images taken about the static scene from different viewpoints, $I_k, k = 0, 1, ..., n-1$. The target is to compute the depth of all pixels in the reference image I_0, that is, find a label map $f : I_0 \to \mathcal{L}$, where $\mathcal{L}$ is a set of labels corresponding to discretized depths in the viewpoint of I_0, encoded by the function $Depth(l), l \in \mathcal{L}$.

For clarity and simplicity of presentation, let I denote any one of the other images since they are treated equivalently. We also assume that $n = 2$ in the following description, while all definition, formulation and conclusions can be easily generalized for more than two input images.

Since the images are calibrated, a warping function $w : I_0 \times \mathcal{L} \to I$ can be defined, which maps a pixel $p \in I_0$ with label l to its corresponding pixel $q \in I$, $w(p, l) = q$. Let the inverse warping function $w^{-1} : I \times \mathcal{L} \to I_0$ map a pixel $q \in I$ to the pixel $p \in I_0$ with label l, $w^{-1}(q, l) = p$.

A warped pixel $w(p_1, l_1)$ occludes another pixel $w(p_2, l_2)$ if and only $w(p_1, l_1) = w(p_2, l_2)$ and $Depth(l_1) < Depth(l_2)$. For multiple pixels warped into the same location in I, only the one with the smallest depth is visible and it occludes all other pixels. Let $O(p, l)$ denote the set of pixels which occlude pixel p with label l in the current label map f,

$$\begin{aligned} O(p,l) = \{q | w(q, f(q)) = w(p, l) \\ \wedge Depth(f(q)) < Depth(l))\} \end{aligned}$$

Given a label map f, a visibility function $V^f : I_0 \to \{0, 1\}$ is defined to return 1 when pixel p is visible after warping and 0 otherwise,

$$V^f(p) = \begin{cases} 1, & O(p, f(p)) = \emptyset \\ 0, & otherwise. \end{cases}$$

The energy function consists of the following two terms,

$$E(f) = E_{data}(f) + E_{smooth}(f) \qquad (1)$$

The data term E_{data} summarizes the pixel-based data term that measures the consistency between the label map and the observations in the images,

$$E_{data} = \sum_{p \in I_0} D(p, f).$$

In most previous methods, the term $D(p, f)$ simply takes a local matching cost function $C(p, f(p))$, such as SSD, SAD or correlation. Our data term differs in that it incorporates the more global visibility constraint. Matching costs are only computed for those pixels visible in other images, and a constant occlusion cost λ_{occ} is assigned for occluded pixels, such as done in dynamic programming [3],

$$D(p, f) = \begin{cases} C(p, f(p)), & V^f(p) = 1 \\ \lambda_{occ}, & V^f(p) = 0. \end{cases}$$

For formulation convenience, it is rewritten as

$$D(p, f) = \lambda_{occ} + V^f(p) \cdot \widetilde{C}(p, f(p) \qquad (2)$$

where $\widetilde{C}(p, f(p)) = C(p, f(p)) - \lambda_{occ}$. The constant λ_{occ} usually takes a value larger than the matching costs observed for correctly matched pixels [8]. In our approach, it is set to the maximum of matching cost due to technical reasons as described in Section 3.2.

The smoothness term $E_{smoooth}$ encodes smoothness prior of the depth map. To illustrate the insensitiveness of our approach to this term, it is simply defined as

$$E_{smooth}(f) = \sum_{(p,q) \in \mathcal{N}} \lambda_{smooth} \cdot S(f(p) - f(q))$$

where $\mathcal{N} = \{(p, q) \mid |p_x - q_x| + |p_y - q_y| = 1\}$ is the 4-connected neighborhood system, $S(\cdot)$ is the simple delta function and λ_{smooth} is a small constant.

3. Energy Function Optimization

This section addresses the minimization of energy function (1). Graph cut can efficiently minimize a function that is defined on binary variables and satisfies certain conditions. This is simply described in Section 3.1. Since a pixel can take multiple labels, previous graph cut stereo methods use α-expansion to compute a strong local minima. In

Section 3.2, we discuss the conditions necessary to apply α-expansion algorithm in our case and show that it is in general infeasible to perform the minimization on all pixels simultaneously using graph cut. Instead, effective approximation techniques are developed which work well in practice. They are elaborated in Section 3.3 and 3.4.

3.1. Graph Cut Algorithm

A theory about energy minimization using graph cut algorithm is established in [11] and simply summarized here.

Let $\{x_1, ..., x_n\}, x_i \in \{0,1\}$, be a set of binary variables. If an energy function consists of a sum of functions of up to three variables,

$$E(x_1,...,x_n) = \sum_i E^i(x_i) + \sum_{i<j} E^{i,j}(x_i,x_j) + \sum_{i<j<k} E^{i,j,k}(x_i,x_j,x_k),$$

it can be minimized by constructing a graph and computing the maximum flow [7], if and only if each term is regular.

Here a single-variable function is always regular. A function of two variables $E(x,y)$ is called regular if it satisfies the following *regularity condition*,

$$E(0,0) + E(1,1) \le E(0,1) + E(1,0) \qquad (3)$$

A function of three variables degenerates to a function of two variables by fixing one variable's value, therefore accounting for totally 6 cases (3 variables and 2 values). The function is called regular if all six two-variable functions are regular.

3.2. α-expansion

Consider the current label map f and a label α, a label map f^α is called within an α-expansion of f if for any pixel $p \in I_0$ either $f^\alpha(p) = f(p)$ or $f^\alpha(p) = \alpha$. Therefore, f^α can be represented by a set of binary variables, $x = \{x_p | p \in I_0, f(p) \neq \alpha\}$[1], such that $f^\alpha(p) = f(p)$ if $x_p = 0$, and $f^\alpha(p) = \alpha$ if $x_p = 1$. Let f^α_x denote the label map represented by x, we can define the following energy function on the binary variables x,

$$E^\alpha(x) = E^\alpha_{data}(x) + E^\alpha_{smooth}(x) \qquad (4)$$

, where $E^\alpha_{data}(x) = E_{data}(f^\alpha_x)$ and $E^\alpha_{smooth}(x) = E_{smooth}(f^\alpha_x)$. In the α-expansion, we compute the variables x that minimizes the energy function (4) using graph cut, and change the label map f to f^α_x. This process is performed over all labels iteratively until convergence. The result proves to be a strong local minima of (1) [4].

It has been shown in [10] that the smoothness term $E^\alpha_{smooth}(x)$ satisfies the regularity condition (3), and the remaining problem is to verify whether the data term $D(p, f^\alpha_x)$ satisfies the regularity condition. Let us assume that $f(p) \neq \alpha$ and $D(p, f^\alpha_x)$ is therefore dependent on x_p. The conclusion when $f(p) = \alpha$ is given later as a special case.

Assume that $V(p)$ reduces to $V_0(p)$ and $V_1(p)$ when x_p takes 0 and 1, respectively,

$$V(p) = (1 - x_p) \cdot V_0(p) + x_p \cdot V_1(p). \qquad (5)$$

Combining (2) and (5) gives the explicit form of data term $D(p, f^\alpha_x)$ as the sum of two terms D_0, D_1 and a constant λ_{occ},

$$\begin{aligned} D(p, f^\alpha_x) = & \lambda_{occ} + x_p \cdot V_1(p) \cdot \widetilde{C}(p,\alpha) \\ & +(1 - x_p) \cdot V_0(p) \cdot \widetilde{C}(p, f(p)) \\ = & \lambda_{occ} + D_1 + D_0. \end{aligned} \qquad (6)$$

Now the problem boils down to finding whether D_0 and D_1 satisfy the regularity condition.

Let q denote a pixel that may occlude p after α-expansion. There are four combinations for x_p and x_q.

- $x_p = 0, x_q = 0$, q occludes p in current label map f and $q \in O(p, f(p))$, denoted as $O_0(p)$;
- $x_p = 0, x_q = 1$, q takes label α and occludes p with its label unchanged. This is only possible when $Depth(\alpha) < Depth(f(p))$, and such q is uniquely determined by $q^* = w^{-1}(w(p, f(p)), \alpha)$;
- $x_p = 1, x_q = 0$, p takes label α and is occluded by q with its label unchanged; $q \in O(p, \alpha)$, denoted as $O_1(p)$;
- $x_p = 1, x_q = 1$, it is impossible for q to occlude p in this case since they have the same label α.

Notice that p will be occluded if any such q occludes it, and it is visible only if all such qs do not occlude it. Summarizing the above four cases gives rise to[2]

$$V_0(p) = \begin{cases} (1 - x_{q^*}) \prod_{q \in O_0(p)} x_q, & \text{if } q^* \text{ exists} \\ \prod_{q \in O_0(p)} x_q, & \text{otherwise} \end{cases} \qquad (7)$$

and

$$V_1(p) = \prod_{q \in O_1(p)} x_q \qquad (8)$$

From (6) and (8), term D_1 is a function of $|O_1(p)| + 1$ variables. It takes $\widetilde{C}(p, \alpha)$ when all variables take 1, and 0 otherwise. When $O_1(p) = \emptyset$, $V_1(p)$ simply vanishes in (6). D_1 becomes a function of one variable x_p and trivially

[1] a pixel whose label is currently α will be a constant, not a variable.

[2] It is possible that $q^* \in O_0$ and $f(q^*) = \alpha$. In this case, there is not corresponding x_{q^*}, the visibility function V_0 is 0 and the term D_0 trivially satisfy the regularity condition. This trivial case is not discussed in the text.

satisfies the regularity condition. When $O_1(p) \neq \emptyset$, since the graph cut algorithm can handle a term of at most three variables in the current state of art [11][3], by checking the regularity condition described in Section 3.1, we have the following conclusion,

Lemma 1 *D_1 can be minimized by graph cut if (i) $O_1(p) = \emptyset$; or (ii) $|O_1(p)| \leq 2$ and $C(p, \alpha) \leq \lambda_{occ}$.*

Similar conclusions can be drawn about term D_0 given by (6) and (7).

Lemma 2 *If q^* exists, D_0 can be minimized by graph cut if $O_0(p) = \emptyset$ and $C(p, f(p)) \leq \lambda_{occ}$.*

Lemma 3 *If q^* does not exist, D_0 can be minimized by graph cut if (i) $O_0(p) = \emptyset$; or (ii) $|O_0(p)| = 1$ and $C(p, f(p)) \geq \lambda_{occ}$.*

Remember that all the above conclusions assume $f(p) \neq \alpha$. When $f(p) = \alpha$, x_p vanishes in (5), $O_0(p) = O_1(p)$ and q^* does not exist. The data term (6) reduces to

$$D(p, f_x^\alpha) = \prod_{q \in O_0(p)} x_q \cdot \widetilde{C}(p, f(p)) + \lambda_{occ}. \qquad (9)$$

Similarly, we have

Lemma 4 *For a pixel p with $f(p) = \alpha$, data term (9) can be minimized by graph cut if (i) $O_0(p) = \emptyset$; or (ii) $|O_0(p)| \leq 3$ and $C(p, \alpha) \leq \lambda_{occ}$.*

There are two problems. The first problem is with respect to the size of $O_{0(1)}(p)$. This is less inherent as it is due to the maximum number of variables in a term manipulable by graph cut [11] and may be relieved with further development in theory. In practice, such conditions are usually satisfied since the scene geometry seldom contains more than two layers in a small area. The second problem is due to the regularity condition (3) and concerns the relation between λ_{occ} and pixel matching cost function $C(p, f(p))$. Unfortunately, Lemma 3 contradicts other cases on this relation and this makes a general solution impossible.

Theorem 1 *Energy function (4) cannot be minimized by graph cut in general.*

Notice that the condition $C(p, f(p)) \geq \lambda_{occ}$ in Lemma 3 is also not reasonable. In the following, we set λ_{occ} to the maximum value of function $C(p, f(p))$. The main difficulty now is how to handle pixel p with $O_0(p) \neq \emptyset$ in Lemma 2 and 3. Fortunately, since those pixels occluded in the current label map usually amount to a small proportion, this makes effective approximation techniques possible, either by ignoring those pixels (Section 3.3) or estimating their data terms instead of accurately evaluating them (Section 3.4).

3.3. Restricted α-expansion

In the restricted α-expansion, those pixels occluded in the current label map are not considered as variables, their labels are not changed and their data terms are also excluded from (4). We are actually minimizing an energy function on a subset of pixels,

$$E^\alpha(\tilde{x}) = \widetilde{E}^\alpha_{data}(\tilde{x}) + E^\alpha_{smooth}(\tilde{x}) \qquad (10)$$

where

$$\tilde{x} = \{x_p | p \in I_0, f(p) \neq \alpha \wedge O_0(p) = \emptyset\}$$

and

$$\widetilde{E}^\alpha_{data}(\tilde{x}) = \sum_{p \in I_0 \wedge O_0(p) = \emptyset} D(p, f_{\tilde{x}}^\alpha)$$

Theorem 2 *Energy function (10) can be minimized by graph cut.*

The proof is straightforward by checking the regularity condition for each term in (10) and it is omitted due to space limitation. The main problem here is that, if a pixel is incorrectly occluded currently, it does not have a chance to change its label in restricted α-expansion even if its true label should be α. This results in a lot of holes in the final depth map, corresponding to those incorrectly occluded pixels (see Figure 1(c)). For most of other visible pixels, however, restricted α-expansion generates correct result. This makes a reliable basis for approximation techniques. One approximation method is described in the following section.

3.4. Approximate α-expansion

As shown in Section 3.2, the data term D_0, D_1 given by (6), (7), (8) (when $f(p) \neq \alpha$) and $D(p, f_x^\alpha)$ (9) (when $f(p) = \alpha$) cannot be minimized by graph cut when they either involve more than three variables or do not satisfy the regularity condition. In the approximate α-expansion, these terms are not accurately evaluated but estimated in a way that they can be minimized by graph cut and the maximum flexibility is retained to make full use of the power of graph cut. Since the number of approximated terms usually account for a small proportion (less than 10% in all experiments), the approximated energy function is close to the true one (4) and the minimization gives satisfactory results in practice. Let such minimized function value be E_{est} and resulting optimal variables be x_{opt}. Evaluating function (4) using x_{opt} gives a value E_{opt}. The approximation error $\frac{|E_{opt} - E_{est}|}{E_{opt}}$ is no more than 1% in all experiments.

Let the probability function $\mathcal{P} : I_0 \to [0, 1]$ measure the probability that pixel p changes its label to α, $\mathcal{P}(p) = Pr(x_p = 1)$. Since restricted α-expansion gives reliable

[3] It is yet unknown that whether this is an upper bound. Consequently, the conditions listed in Lemma 1–4 are sufficient and it is unknown whether they are necessary.

results on visible pixels, these pixels are assigned a high certainty ρ (0.9 in the experiment) according to the result of restricted α-expansion,

$$\mathcal{P}(p) = \begin{cases} \rho, & f_{\tilde{x}}^{\alpha}(p) = \alpha \\ 1-\rho, & f_{\tilde{x}}^{\alpha}(p) \neq \alpha \end{cases}$$

, where $f_{\tilde{x}}^{\alpha}$ is the label map generated by restricted α-expansion.

For those occluded pixels, their probabilities are heuristically estimated as

$$\mathcal{P}(p) = \frac{D(p, f_{\tilde{x}}^{\alpha})}{D(p, f_{\tilde{x}}^{\alpha}) + D(p, f_{\tilde{x}}^{\alpha}(p))}$$

, where $f_{\tilde{x}}^{\alpha}(p)$ denotes a label map by changing the label of pixel p to α in $f_{\tilde{x}}^{\alpha}$.

Those terms that cannot be manipulated by graph cut are estimated as follows. In Lemma 1, when $|O_1(p)| \geq 3$, the term D_1 is approximated as a function of three variables that satisfies the regularity condition,

$$\begin{aligned} \widetilde{D}_1(x_{q_1}, x_{q_2}, x_{q_3}) &= x_{q_1}x_{q_2}x_{q_3} \cdot (x_p \textstyle\prod_{q \in \widetilde{O}_1(p)} x_q)\tilde{C}(p,\alpha) \\ &\simeq x_{q_1}x_{q_2}x_{q_3} \cdot (\mathcal{P}(p) \textstyle\prod_{q \in \widetilde{O}_1(p)} \mathcal{P}(q))\tilde{C}(p,\alpha) \end{aligned} \tag{11}$$

, where q_1, q_2, q_3 are three pixels in $O_1(p)$ with the largest depths and $\widetilde{O}_1(p) = O_1(p) - \{q_1, q_2, q_3\}$. Dependency between variables is ignored here.

In Lemma 2, when $O_0(p) \neq \emptyset$, the term D_0 is approximated as a function of two variables x_p and x_{q^*},

$$\begin{aligned} \widetilde{D}_0(x_p, x_{q^*}) &= (1-x_p)(1-x_{q^*})(\textstyle\prod_{q \in O_0(p)} x_q)\tilde{C}(p, f(p)) \\ &\simeq (1-x_p)(1-x_{q^*})(\textstyle\prod_{q \in O_0(p)} \mathcal{P}(q))\tilde{C}(p, f(p)). \end{aligned} \tag{12}$$

In Lemma 3, when $O_0(p) \neq \emptyset$, the term D_0 is approximated as a function of one variable x_p and this is a special case of (12) where x_{q^*} vanishes.

In Lemma 4, when $|O_0(p)| > 3$, the term in (9) is approximated as a function of three variables and this is a special case of (11) where x_p vanishes.

3.5. Minimization

The minimization algorithm iterates over all possible labels. The labels are in descending order according to their frequencies in a label map computed by a simple local method with a 3×3 window. This strategy slightly improves the result than using randomly ordered labels. For each label α, restricted α-expansion is performed at first and approximate α-expansion is done afterwards. The label map f is then changed accordingly. Since the energy function is approximated, convergence cannot be determined and the algorithm is terminated after two iterations.

For multi-view stereo ($n > 2$), the data term in energy function (1) becomes

$$E_{data} = \sum_{i=1}^{n-1} \sum_{p \in I_0} D_i(p, f) = \sum_{i=1}^{n-1} \sum_{p \in I_0} (\lambda_{occ} + V_i^f(p) \cdot \widetilde{C}_i(p, f(p)).$$

All techniques described above can be applied similarly. It is worth noting that those pixels ignored in the restricted α-expansion are the union of the pixels occluded in any one of other views.

4. Experiments

Binocular stereo The validity of the occlusion handling of proposed approach is demonstrated using the binocular stereo data Map in Figure 1. The scene consists of two large slanted planar surfaces. Severe occlusion is present on the left of front surface in the reference image and shown as red in Figure 1(b). As discussed in Section 3.3, the restricted α-expansion generates a lot of holes corresponding to incorrectly occluded pixels (Figure 1(c)), but is accurate for those visible pixels. Consequently, the energy function can be effectively estimated in the approximate α-expansion and those holes are removed in the final result (Figure 1(d)). Our result compares favorably with results obtained using other global energy minimization methods, as shown in second row in Figure 1.

The results on Tsukuba data is demonstrated in Figure 2. The scene geometry is complex and does not satisfy the ordering constraint, noticing the long and thin lamp pole in Figure 2(a). This data consists of five images with cameras located as a cross. The (middle,right) pair is used for the binocular stereo experiment. For comparison on the same platform, the algorithms in [4] (short for GC) and [10] (short for Multicam GC) are implemented on our own. The results using Multicam GC and our method are shown in Figure 2(b) and Figure 2(c), respectively. Since the area between the lamp poles contains few texture information, those pixels are assigned the front depth incorrectly. This ambiguity is resolved when more images are used.

Multi-view stereo In the multi-view stereo experiment for Tsukuba data, the middle image is selected as the reference image. The results using GC, MultiCam GC and our method are shown in Figure 2(e), (f) and (g), respectively. Figure 2(h) is our result without using E_{smooth} in the energy function. It shows that the occlusion computing process has a global effect and encourages depth smoothness along epipolar lines, therefore reducing the dependency on the smoothness term.

The running time and error statistics on Tsukuba data are reported in Table 2. MultiCam GC and our method generate results with similar quality, while GC gives noisier results near the depth discontinuities due to the absence of explicit occlusion handling. It is worth noting that the error rate e_{all} for GC using 5 views is even higher than that when using 2 views, that is, the information from more views is actually harmful. This may reveal the importance of occlusion handling in multi-view stereo.

Figure 3 shows the results on Garden data consisting of 11 images. The result of GC (Figure 3(c)) is noisier and obvious fore-ground fattening artifacts near the boundary of the frontal tree can be observed. The second and third rows are the results by our method and MultiCam GC, respectively. The comparable quality in different cases demonstrates that MultiCam GC does not gain an advantage over our method for multi-view stereo. See more discussions in the caption of Figure 3.

Our method is not obviously advantageous over MultiCam GC in running performance for Tsukuba and Garden data. This is mainly because our algorithm currently runs two α-expansions for each label, therefore twice the number of iterations. In addition, since the image sizes are relatively small (384×288 and 344×240, respectively), the saving in the graph cut minimization is not very significant.

For more images with larger size, more running resource is consumed in the graph cut minimization which simplified in proposed approach. Therefore our approach becomes more advantageous by trading little accuracy loss for large running performance improvement. This is demonstrated in Figure 4 on the very challenging Samsung data consisting of 11 images with dimension 640×486. In MultiCam GC, if we use the energy function consisting of all two-view functions, only 7 views can be manipulated on our pc (750 MB memory consumed with maximum 1GB memory). Using 11 views simultaneously will consume 1.8 GB memory, while our approach uses 310 MB.

Parameter setting Our algorithm involves a few parameters and they are fixed in all the experiments, $\lambda_{occ} = 10$ and $\lambda_{smooth} = 3$. For the matching cost function C (truncated by λ_{occ}), we use the pixel dissimilarity measure in [2] that proves to be insensitive to the image sampling noise. In our implementation of GC and MultiCam GC, the parameter λ_{smooth} is allowed to vary to obtain the best result, either by comparison to ground truth or visual inspection when ground truth is not available. GC method usually needs a larger λ_{smooth} to achieve results with similar quality.

5. Discussions

The proposed approach is shown to be comparable with the state of the art. Compared with the multi-view graph cut approach [10], our formulation significantly reduces the size of the graph by using much fewer number of pixels in the occlusion reasoning. Consequently, it computes one depth map with similar quality more efficiently. This extends the applicability of using graph cut in stereo.

There leaves a lot of room for improvement. Current approximate minimization techniques are somewhat heuristic. It is unclear that to what extent the energy function (4) can be accurately minimized and this may deserve further investigation. In the current implementation, actually two α-expansions are performed for one label. Most of the computation in the approximate α-expansion is redundant and this nearly doubles the running time. If more effective approximation method can be developed to run only one α-expansion for each label, this can hopefully reduce the running time nearly by half.

Acknowledgments

This work is supported by the Hong Kong RGC grant HKUST 6182/04E and 6188/02E.

References

[1] http://www.middlebury.edu/stereo/.

[2] S. Birchfield and C. Tomasi. A pixel dissimilarity measure that is insensitive to image sampling. IEEE *Trans. on Pattern Analysis and Machine Intelligence*, 20:401–406, 1998.

[3] A. F. Bobick and S. S. Intille. Large occlusion stereo. *International Journal of Computer Vision*, 33(3):181–200, 1999.

[4] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. In *Proceedings of the 7th Int. Conf. on Computer Vision*, pages 377–384, September 1999.

[5] M. Z. Brown, D. Burschka, and G. D. Hager. Advances in computational stereo. IEEE *Trans. on Pattern Analysis and Machine Intelligence*, 25(8):993–1008, August 2003.

[6] A. Criminisi, J. Shotton, A. Blake, and P. Torr. Gaze manipulation for one-to-one teleconferencing. In *Proceedings of the 9th Int. Conf. on Computer Vision, Nice, France*, 2003.

[7] L. Ford and D. Fulkerson. *Flows in Networks*. Princeton University Press, 1962.

[8] S. B. Kang, R. Szeliski, and J. X. Chai. Handling occlusions in dense multi-view stereo. In *Proceedings of the Conf. on Computer Vision and Pattern Recognition*, 2001.

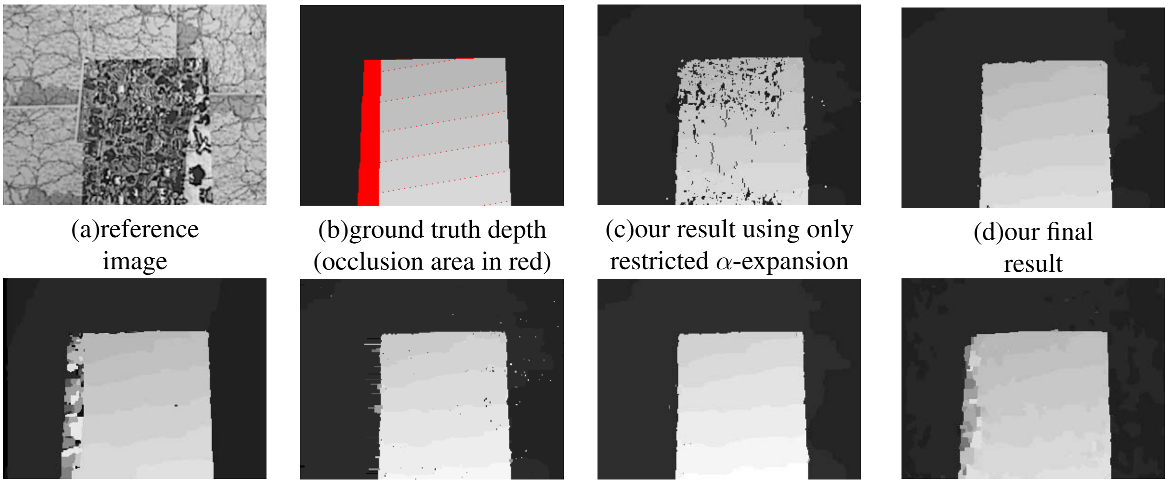
[9] V. Kolmogorov and R. Zabih. Computing visual correspondence with occlusions via graph cuts. In *Proceedings of the 8th Int. Conf. on Computer Vision*, volume 2, pages 508–515, 2001.

[10] V. Kolmogorov and R. Zabih. Multi-camera scene reconstruction via graph cuts. In *Proceedings of the 7th European Conf. on Computer Vision*, 2002.

[11] V. Kolmogorov and R. Zabih. What energy functions can be minimized via graph cuts? In *Proceedings of the 7th European Conf. on Computer Vision*, 2002.

[12] M. Lhuillier and L. Quan. Match propogation for image-based modeling and rendering. IEEE *Trans. on Pattern Analysis and Machine Intelligence*, vol 24:pages 1140–1146.

[13] J. Sun, H. Y. Shum, and N. N. Zheng. Stereo matching using belief propagation. In *Proceedings of the 7th European Conf. on Computer Vision*, 2002.



(a)reference image

(b)ground truth depth (occlusion area in red)

(c)our result using only restricted α-expansion

(d)our final result

Graph cuts [4]

GC+occl [9]

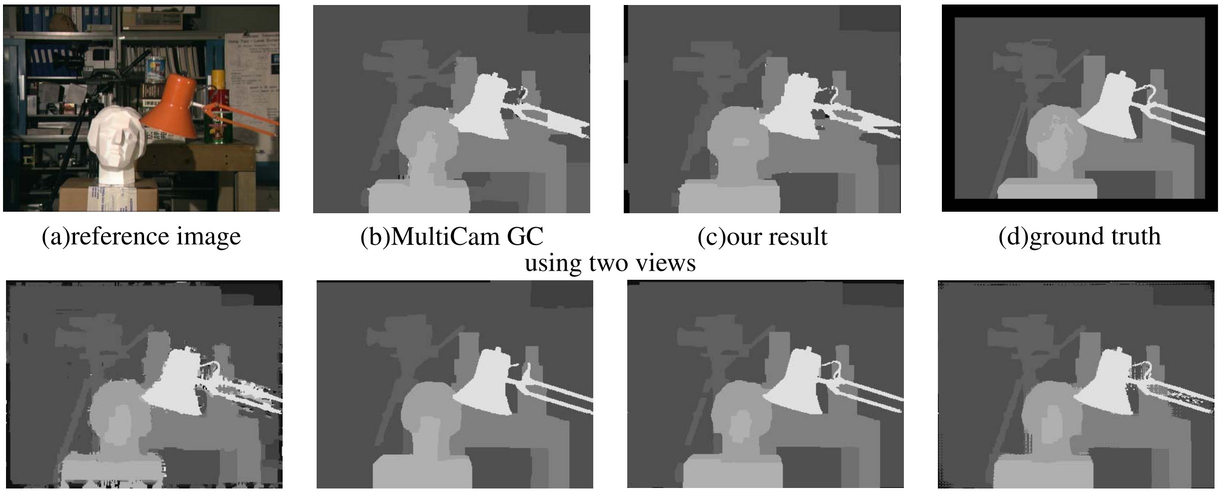
MultiCam GC [10]

Belief prop [13]

Figure 1. Result for Map data. The second row shows the result in other global energy minimization methods obtained from middlebury database [1].

using 2 views	time	e_{occ}	e_{disc}	e_{all}	using 5 views	time	e_{occ}	e_{disc}	e_{all}
GC	7	88.5	11.2	4.3	GC	8.6	58.3	18.6	5.46
MultiCam GC	16	31.5	7.1	2.2	MultiCam GC	40	12.4	6.64	1.28
our method	16	37.3	13.4	2.68	our method	38	16.7	5.2	1.30

Table 2. Running time(in seconds) and error statistics(%) for Tsukuba data. A label differing from the true value by more than one level is considered erroneous. The error rate is computed over occluded areas (e_{occ}), depth discontinuity areas (e_{disc}) and the entire image (e_{all}), respectively.



(a)reference image

(b)MultiCam GC

(c)our result

(d)ground truth

using two views

(e)GC

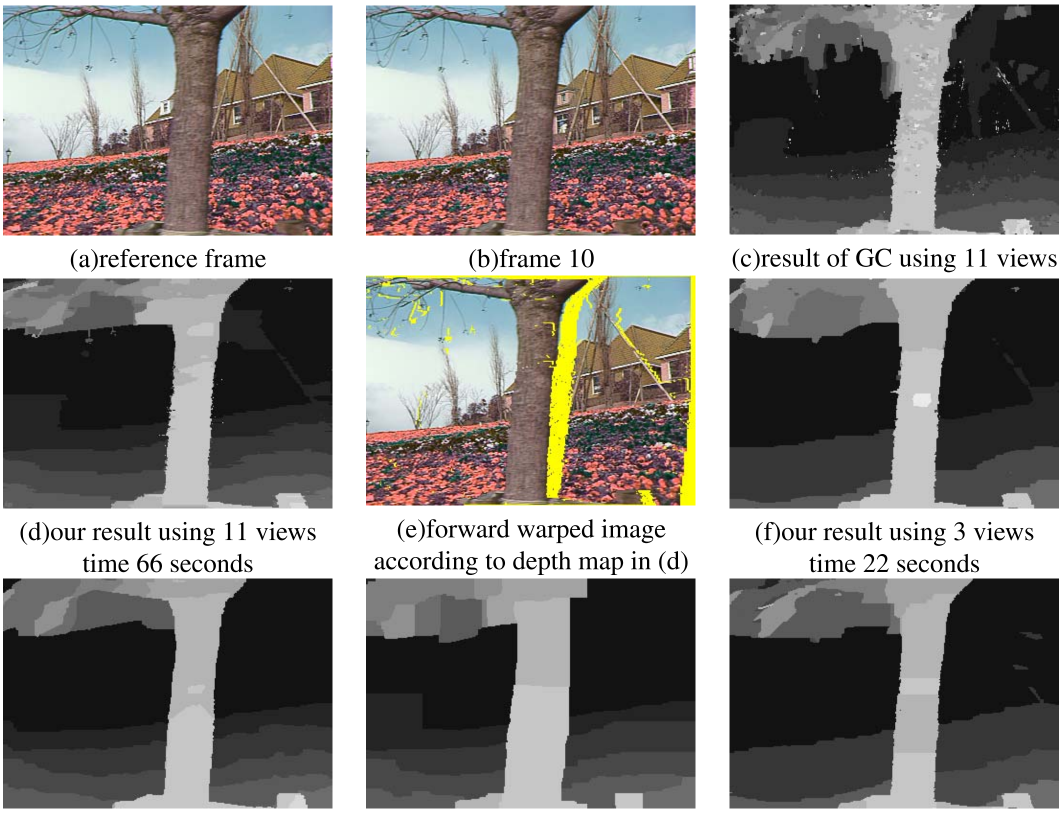
(f)MultiCam GC

(g)our result

(h)our result, no E_{smooth}

using five views

Figure 2. Result for Tsukuba data. The results using GC [4] and MultiCam GC [10] are obtained in our implementation using best smoothness parameters. The running time and error statistics are reported in Table 2. Result in (h) demonstrates the effectiveness of occlusion handling without using smoothness term in the energy function.



(a)reference frame

(b)frame 10

(c)result of GC using 11 views

(d)our result using 11 views
time 66 seconds

(e)forward warped image
according to depth map in (d)

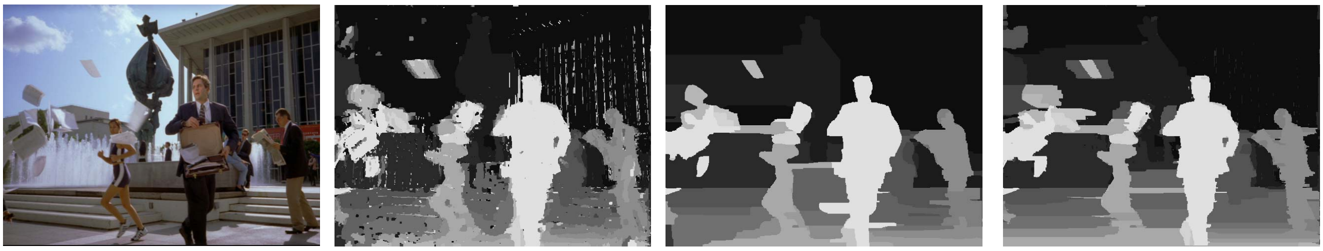
(f)our result using 3 views
time 22 seconds

(g) $\{(i, j)|i < j\}$
time 713 seconds

(h) $\{(ref, i)|i \neq ref\}$
time 155 seconds

(i) $\{(i, i + 1)\}$
time 150 seconds

Figure 3. Result for Garden data. The third row shows the results of MultiCam GC using different sets of two-view energy functions. When all possible image pairs are used (g), the result is comparable to (d) and the running time is proportional ($713/66 \simeq 11$). If only the image pairs containing reference image are used, the depth map of reference image is similar to (g), but other depth maps are worse ((h) is the depth map for frame 10 and fore-grounding fattening is obvious) due to the asymmetric setting. If only consecutive image pairs are considered (i), the result is comparable to (f) which uses 3 consecutive views. This reveals that MultiCam GC method does not gain an advantage over our approach for multi-view stereo in terms of either quality or running performance.



(a)reference image

(b)GC

(c)MultiCam GC

(d)our result

Figure 4. Result for Samsung data. All results are obtained using 7 views. The running time for MultiCam GC and our approach is 1200 seconds and 262 seconds, respectively. Our result has similar quality as in MultiCam GC by inspection, and is smoother than the result by GC.