Acceleration Strategies in Generalized Belief Propagation

Shengyong Chen, *Senior Member, IEEE*, and Zhongjie Wang

***Abstract*—Generalized belief propagation is a popular algorithm to perform inference on large-scale Markov random fields (MRFs) networks. This paper proposes the method of accelerated generalized belief propagation with three strategies to reduce the computational effort. First, a min-sum messaging scheme and a caching technique are used to improve the accessibility. Second, a direction set method is used to reduce the complexity of computing clique messages from quartic to cubic. Finally, a coarse-to-fine hierarchical state-space reduction method is presented to decrease redundant states. The results show that a combination of these strategies can greatly accelerate the inference process in large-scale MRFs. For common stereo matching, it results in a speed-up of about 200 times.**

***Index Terms*—Accelerated generalized belief propagation (AGBP), computer vision, high order, inference, Markov random fields (MRFs), pattern analysis.**

I. Introduction

MARKOV RANDOM FIELDs (MRFs) is a widely used graphical model in various disciplines such as statistical learning, pattern analysis, coding and decoding, and image understanding. Many applications in these areas can be found in industrial informatics [1]–[3]. The inference of MRFs has received profound research [4], [5]. Among heterogeneous approaches, belief propagation (BP) is one of the most superior methods [6]. In order to acquire marginal posterior probabilities, it uses a message-passing strategy to iteratively update beliefs and messages for every variable. In graphical models without loops, this approach had proven to provide the exact marginal posterior probabilities, whereas in graphical models with massive loops, as in the grid-like MRFs, the convergence problem becomes more difficult. To accelerate convergence, many approaches have been proposed and some significant progress has been achieved. However, due to the slow convergence of BP in graphical models with loops, the interest of BP has decreased. On the other hand, generalized belief propagation (GBP) proposed by Yedidia *et al.* [7] with its better convergence property compared to BP has recently received a large attention.

GBP can be considered as an extension of BP. It is also an instance of cluster variation methods. As described in the literature, BP can only converge to a stationary point of Bethe free energy, while GBP can converge to a stationary point of Kikuchi free energy which is a more stable approximation than Bethe free energy [8], which leads to an improved convergence property. Despite its good convergence, GBP requires a large computational effort. Without any optimization, the canonical version of GBP takes a quartic temporal complexity, while the fast version of BP in [9] reaches linear complexity. This severely restricts its applicability in the inference of low-order graphical models and, obviously, prevents the application of GBP for more complicated problems [10], [11].

This paper proposes an approach to accelerate GBP, which is called accelerated generalized belief propagation (AGBP) in the following. This paper focuses on the problem of inference in large-scale grid-like MRFs. The term "large-scale" referred here to the large number of variables but also to the large state space for each node. To make GBP more efficient in this kind of graphical models, Petersen *et al.* [11] propose two acceleration methods to improve the performance of GBP. The first one is a caching strategy. By caching some precomputed variables, many redundant computations can be avoided. The first acceleration method proposed in this paper adopts this method and achieves a similar acceleration rate. The second acceleration method proposed in [11] is to convert a searching problem on a grid into a linear searching problem. Since the converted linear searching problem should be monotonous, such a method is probably going into a local minimum and the acceleration rate is largely influenced by the size of the square. Compared with previous approaches, the proposed AGBP promises a great acceleration. Strategies are explained in this paper. First, a min-sum messaging scheme and a caching technique are used to improve the accessibility. Second, a direction set method is introduced into the pairwise message computation stage which decreases the temporal complexity from quartic to cubic. Finally, a strategy of hierarchical state-space reduction is proposed to significantly reduce the redundant states in every hierarchical level which can further decrease the quantity of computations.

II. Problem Formulation

This paper takes the example of GBP formulation for a stereo vision system where stereo matching is an inference problem in MRFs for depth map estimation [12], [13]. This can be posed as an energy minimization problem. The corresponding energy function used here is the most conspicuous one which is defined as

$$\mathrm{E}(f) = \sum_{p \in \mathbf{P}} \mathrm{D}(f_p) + \sum_{(p,q) \in \mathbf{N}} \mathrm{V}(f_p, f_q) \qquad (1)$$

Manuscript received March 15, 2011; revised August 25, 2011; accepted September 25, 2011. Date of publication October 24, 2011; date of current version January 20, 2012. This work was supported in part by the National Natural Science Foundation of China and Microsoft Research Asia under Grant 61173096, Grant 60870002, and Grant R1110679. Paper no. TII-11-106.

S. Chen is with the College of Computer Science and Technology, Zhejiang University of Technology, 310023 Hangzhou, China (e-mail: sy@ieee.org).

Z. Wang is with the International Max Planck Research School for Computer Science, 66123 Saarbruecken, Germany (e-mail: zwang@mpi-inf.mpg.de).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TII.2011.2172449

where $\mathbf{N}$ is the four-connected neighborhood set, D denotes the data cost for assigning a label f_p to a pixel p, and V denotes the pairwise cost for labeling between two neighbor pixels p and q. The pairwise cost can induce close pixels to have same labels, which makes the result smooth.

The data cost is calculated by a truncated linear model, and is defined as

$$\mathrm{D}(f_p) = \lambda \cdot \min\left(\sqrt{\sum_{c\in\{L,a,b\}} \left(\mathbf{I}_c^L(p) - \mathbf{I}_c^R(p - f_p)\right)^2}, T\right) \tag{2}$$

where λ is the cost weight which determines the portion of energy that data cost possesses in the whole energy, T represents the truncating value, and $\mathbf{I}_c^L(p)$ and $\mathbf{I}_c^R(p)$ represent the intensity values of pixel p of channel c in the left and right images, respectively. It can improve the final results to some degree as observed from our practical experience.

The pairwise cost is defined as

$$\mathrm{V}(f_p, f_q) = \min\left(|f_p - f_q|, K\right) \tag{3}$$

where K is the truncating value. The pairwise cost smooths the result but also preserves discontinuity, since it can prevent the edges of objects from over smoothing.

The energy function defined in (1) can be considered as a description of the scene. Here, an assumption is made that the minimum of (1) is always matched with the correct scene structure, which is represented by the depth information in stereo matching. In the literature, to make the condition assumption more accurate, a rather complex energy function should be employed. However, to simplify the presentation and to be consistent and comparable with other methods, the dualistic energy function as (1) is used in this paper.

III. Message Passing in AGBP

A. Edge and Cluster Messaging

GBP can be considered as a method of Kikuchi free energy approximation. It allows, in general, an arbitrary number of variables to be a clique and includes the clique information into the whole message-passing process, while BP only allows site-to-site message passing. Since additional sources of information, such as the clique information, are involved in the message-passing process, the search capability is improved for locating the minimum of the energy function.

The approach introduced in [8] comprises two kinds of regions, i.e., single site region and double sites region, and the related messages are named edge message and cluster message, respectively. The message update rules are defined in (4) and (5) shown at the bottom of the page. Illustrations of the message-passing process can be seen in Figs. 1 and 2, where ϕ_x is the local evidence of variable x, and φ_{xy} is the mutual evidence of variable x and y. Equation (4) describes the edge message sending from a specific site s to site u. Equation (5) describes the cluster message sending from sites s and t to sites u and v.

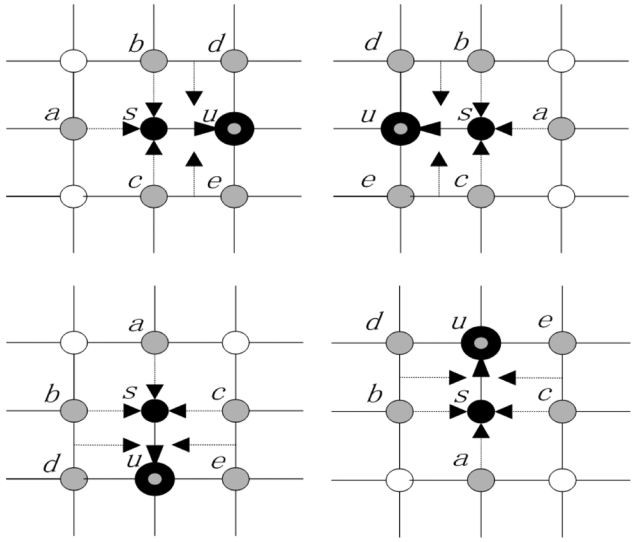


Fig. 1. Edge messages passing from site s to four neighbors.

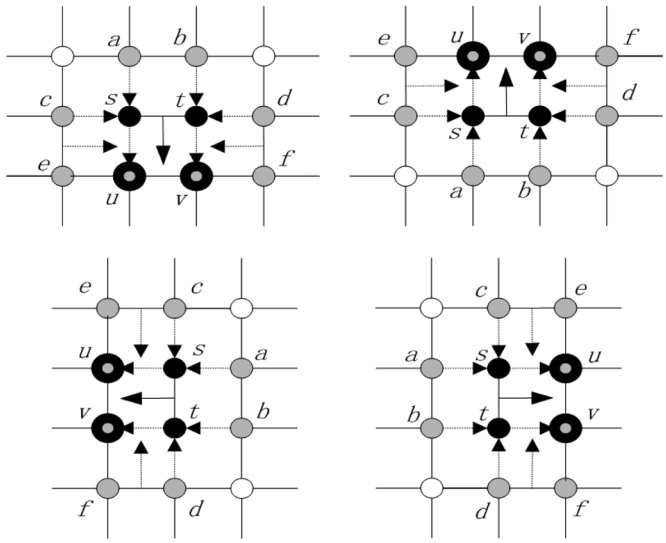


Fig. 2. Cluster messages passing from s, t to u, v.

B. Min-Sum Messaging

Considering the propagation efficiency, this paper proposes an adaptive strategy to achieve improved performance. First,

$$m_{s\to u}(x_u) = \max_{x_s}(\phi_s \varphi_{su} m_{a\to s} m_{b\to s} m_{c\to s} m_{bd\to su} m_{ce\to su}) \tag{4}$$

$$m_{st\to uv}(x_u, x_v) = \frac{\max_{x_s,x_t}(\phi_s \phi_t \varphi_{st} \varphi_{su} \varphi_{tv} m_{a\to s} m_{c\to s} m_{b\to t} m_{d\to t} m_{ab\to st} m_{ce\to su} m_{df\to tv})}{m_{s\to u} m_{t\to v}} \tag{5}$$

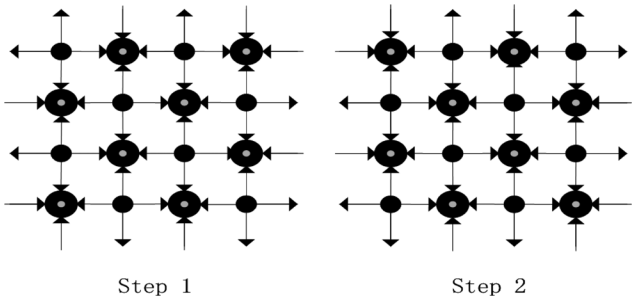


Fig. 3. Chessboard passing for edge messages.

this paper replaces the message-passing scheme in the canonical GBP with a min-sum representation based on (4) and (5), and then a negative logarithm operation is applied. Second, to reduce the computational effort for messages, an analog caching method like [11] but defined in the min-sum scheme is employed. The new corresponding message-passing rules for the energy function are defined as

$$m_{s\to u}^{t}(x_u) = \min_{x_s} \left(P_s(x_s) + Q_{su}(x_s, x_u)\right) \tag{6}$$

where

$$P_s(x_s) = \phi_s + \sum_{i=\{a,b,c\}} m_{i\to s}^{t-1}(x_s) \tag{7}$$

and

$$Q_{su}(x_s,x_u) = \varphi_{su} + \sum_{i=\{bd,ce\}} m_{i\to su}^{t-1}(x_s,x_u) \tag{8}$$

$$m_{st\to uv}^{t}(x_u,x_v) = \min_{x_s,x_t}\left(Q'_{su}(x_s,x_u) + Q'_{tv}(x_t,x_v) + Q'_{st}(x_s,x_t)\right) - m_{s\to u}^{t-1} - m_{t\to v}^{t-1} \tag{9}$$

where

$$Q'_{su} = \phi_s + \varphi_{su} + m_{a\to s}^{t-1}(x_s) + m_{c\to s}^{t-1}(x_s) + m_{ce\to su}^{t-1}(x_s,x_u) \tag{10}$$

$$Q'_{tv} = \phi_t + \varphi_{tv} + m_{b\to t}^{t-1}(x_t) + m_{d\to t}^{t-1}(x_t) + m_{df\to tv}^{t-1}(x_t,x_v) \tag{11}$$

and

$$Q'_{st} = \varphi_{st} + m_{ab\to st}^{t-1}(x_s,x_t) \tag{12}$$

where t is the number of iterations and $P_s(x_s)$, Q'_{su} and Q'_{tv} can be treated as cache variables.

The chessboard passing strategy [9] is applied when the message-passing process is started. Since GBP has two kinds of messages to update, an extended chessboard passing strategy which updates edge messages and cluster messages in turn is proposed. A diagram of such a process is shown in Figs. 3 and 4. Arrows show the directions of the message passing. The process of message passing is separated into two steps. Small solid nodes are the active nodes which are passing messages to their neighbor nodes. Fig. 4 illustrates that clique message passing consists of two parts: horizontal passing and vertical passing.

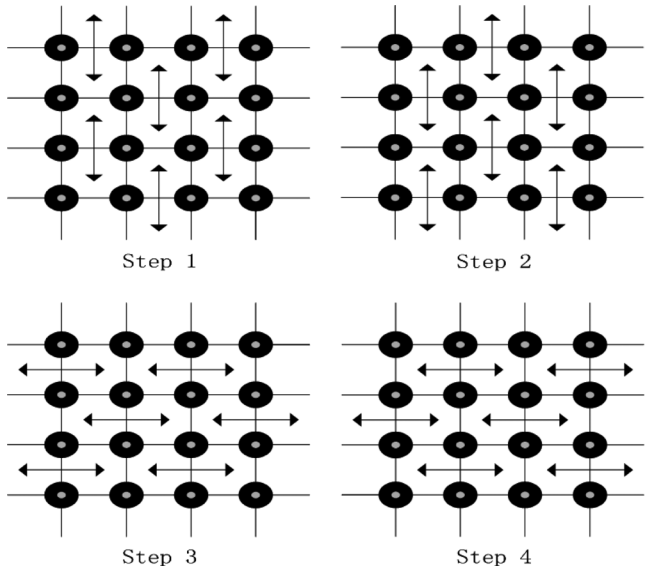


Fig. 4. Chessboard passing for clique messages.

After several iterations, the final result is obtained by calculating the belief of each variable. It can be defined as

$$B_s^t(x_s) = \sum_{i=\{a,b,c,d\}} m_{i\to s}^{t}(x_s). \tag{13}$$

IV. Complexity Reduction

In order to further reduce the complexity of AGBP, this paper introduces two additional strategies for decreasing the exponent number.

1) Applying direction set method to reduce the computation complexity of cluster message from $O(n^4)$ to $O(n^3)$.
2) Using a hierarchical framework and combining with the state-space reduction strategy to significantly squeeze the search space, in other words, to decrease n based on the complexity analysis.

First of all, the first strategy and the second strategy will be explained shortly. From (9), when x_u and x_v are given, the temporal complexity to compute a specific item in the cluster message is $O(n^2)$, where n is the dimension of the state space. The majority of the computations is contributed from the first term in the equation, which can be regarded as finding the minimum value in a square lattice where the two axes are s and t, respectively. If all the elements in the lattice are traversed, the temporal complexity is $O(n^2)$. Petersen *et al.* [11] proposed a method to reduce the search space, but it relies heavily on the traverse order. The method proposed in this study is more straightforward. When the direction set method is applied, the temporal complexity becomes $O(n)$, and the total complexity for computing the cluster message is reduced to $O(n^3)$.

The direction set method adopted here is also called Powell's method [14]. It decomposes an *N*-dimensional (N-D) search problem into several 1-D search processes. Take an example in 2-D lattice where a site P has a random initial position and its two orthogonal directions are given. First, P moves to the extreme value position which is found by searching along the first direction among the two initial directions. Second, P moves to another extreme value position by searching along the second

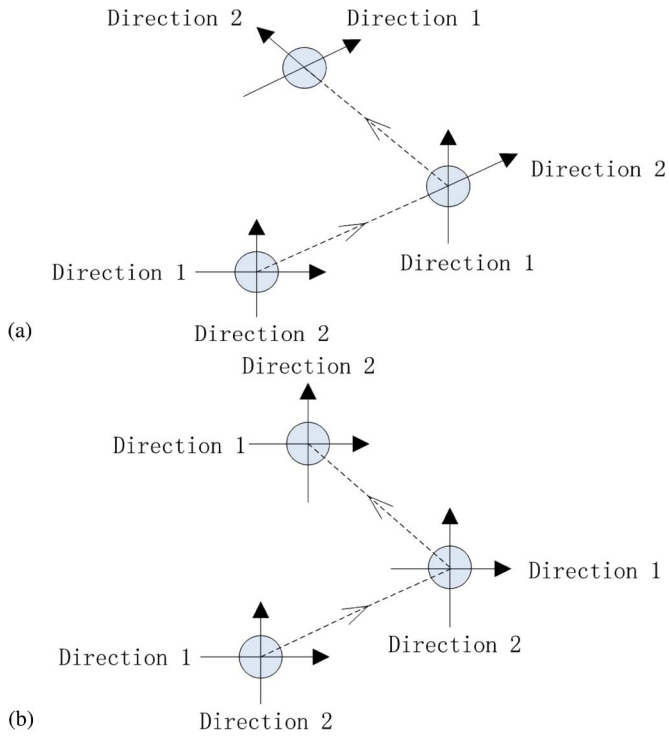


Fig. 5. Two ways of the direction set method. (a) General process of the 2-D direction set method. (b) General process of the static direction set method.

direction given by initialization. Third, the first direction is substituted by the second which is set to be a new direction determined by the initial position and the final position after two rounds of searching. Meanwhile, the final position is set to be the new initial position. The three steps are performed in an iterative way until P no longer moves. Fig. 5(a) illustrates this approach.

The general idea of the direction set is challenging because the two directions will become similar in some cases. Since the two arrays of search positions determined by the two directions also become similar, the search capability in this iteration will be reduced and the process is in a high risk of unexpected ceasing. On the other hand, in practice, it is hard to search along an arbitrary direction where it requires more computation to determine which sites are occupied. The two directions are set to be static and parallel along each axis [see Fig. 5(b)]. This setting not only keeps the orthogonality condition from the beginning to the end, but also makes the implementation easier. Fig. 6 illustrates one iteration in the general process of computing the first part in (9).

Generally, it is more efficient to place the initial position close to the extreme value position. To place it near the minimum value position, an assumption is made that the combination of the independent minimum value positions of s and t is close to the actual minimum value position. In this sense, the initial position (s_{ini}, t_{ini}) can be given as follows:

$$s_{ini} = \arg\min_{s} \{Q'_{su}(x_s, x_u)\} \tag{14}$$

$$t_{ini} = \arg\min_{t} \{Q'_{tv}(x_t, x_v)\} \tag{15}$$

where $Q'_{su}(x_s, x_u)$ and $Q'_{tv}(x_t, x_v)$ are defined in (10) and (11), which are the independent terms about s and t in (9), respectively.

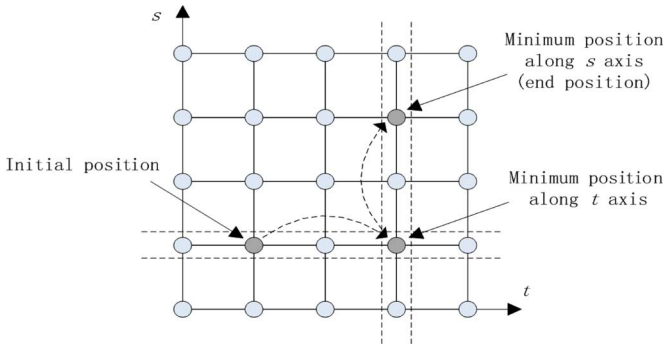


Fig. 6. One iteration in the static direction set method.

Through this optimization, the number of accessed positions is decreased from n^2 to $2kn$, where k is the number of iterations. In our experiments, it is often set as 2 or 3. n is the state space. Since the comparison operation occupies the most computation time, the general complexity becomes $2kn^3$, while the complexity of brute force search is n^4. The efficiency rate is $n/(2k)$. When n is increased, the rate of computation time is also increased.

V. Hierarchical State-Space Reduction

In this section, a hierarchical method, called hierarchical state-space reduction, is introduced to reduce irrelevant states within a step-by-step reduction approach. The hierarchical pyramid is constructed similarly as in [8]. The approach collects every four sites to comprise a new site in the coarser level and takes the data cost of the up-left site as the newly replaced site.

Hierarchical structure has an inherent advantage for message-based algorithms, e.g., BP and GBP. Message passing in a coarser level means a longer range than that in a finer level. In the proposed approach, the information of a variable can be sent to the variables away from it, not being restricted only to its neighbors. It can largely encourage the information flowing in the whole graph and make the system converge more quickly. Because message passing is processed at each level separately, the information about the states which are far away from the true state does not much influence the final result. Therefore, the states far away from the true state can be removed, which can also save a large quantity of computations. However, another question might be aroused that the true state could be accidentally smoothed away at a coarser level. To avoid this problem, this paper proposes a state selection approach to selecting a number of most possible states at a finer level.

First, a minimum number of states that each level should maintain is given. The purpose of this setting is to prevent the states from being reduced so fast that the optimization might be over restricted to a small state space and local convergence. This strategy can maintain a space large enough to provide sufficient transferring information and also keep the computation cost at a low level. Furthermore, it is difficult to determine the balance between accuracy and computation cost. Here, to be general, a constant reduction method is proposed by the following definition

$$\begin{cases} \boldsymbol{\Gamma}(L) = d \\ \boldsymbol{\Gamma}(i) = \boldsymbol{\Gamma}(i-1) - \eta \end{cases} \tag{16}$$

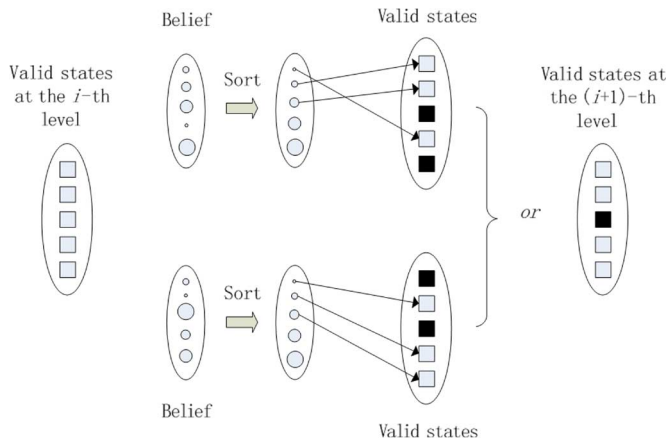


Fig. 7. State selection process. White or black squares represent valid or invalid states. Bigger circles in the belief vector and local evidence vector stand for lower possibility to be the true state. In this example, the valid states of the current site at the ith level is 5, and $\Gamma\ (i+1)$ is 3.

where $\Gamma(i)$ is the minimum number of states at the ith level, d is the number of states at the coarsest level, L is the number of hierarchical levels, and η is a constant of reduction rate. These parameters are mainly related with the searching range, image resolution, and stereo setup. We may easily find a set of empirical numbers from practical experience for a kind of applications.

After setting the minimum numbers of states at each level, a two-way strategy is used to select the actual amount of states for every variable. First, when the energy of the ith level comes to converge, beliefs are computed and sorted, and then $\Gamma(i)$ states with highest score are chosen to be valid. Second, sort the local evidence and also choose a number of the leading states to be valid. Third, two validation results are combined using the logical "OR" operation. This process is shown in Fig. 7. In the ith level, all five states are valid. Through two kinds of state selection process, the third state is set to be invalid in the $(i+1)$th level.

This state selection method has several merits. First, the best states selected by beliefs can be passed down to next level as a result from the coarser level. Second, some wrongly invalidated states are reserved according to their superior local evidences.

After setting the valid states in the next level, the related messages should also be passed over. Since the number of valid states is changed, the item related to the invalid states in messages should be abandoned so that the inherited messages for each variable at the finer level are different. This inheriting strategy is different from that in [9]. In this proposed method, a message inherited from a coarser level reaches two sites away from it at a finer level. It can increase the restriction length from the coarser level and help useful information flowing. The corresponding message inheriting processes are illustrated in Figs. 8 and 9. The main motivation of this approach is to pass a message to all the related variables in a finer level, and thus, it can avoid information loss by the greatest extent. In other words, make the influence from the result of the coarser level to the optimization of the next level the furthest.

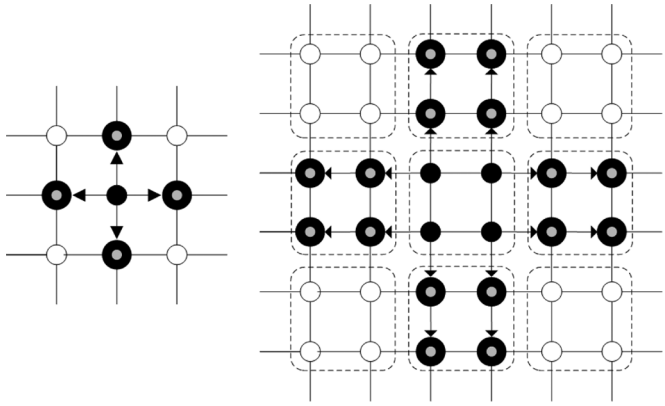


Fig. 8. Edge message inheritance. (Left) Coarse level. (Right) Fine level.

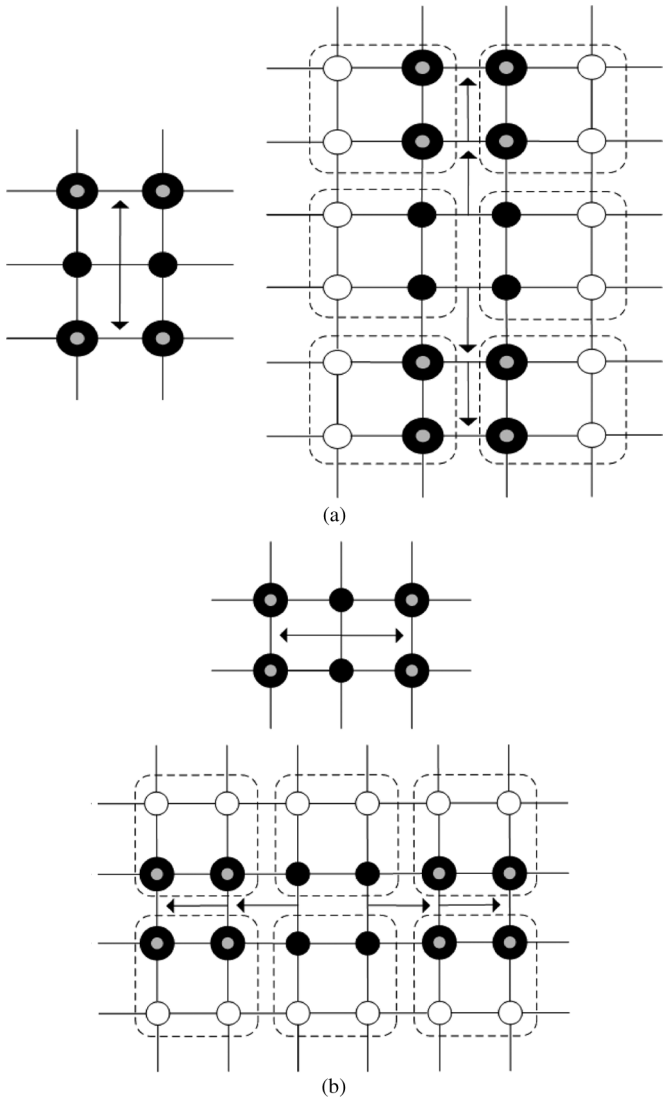


Fig. 9. Clique message inheritance. (a) Vertical clique message inheritance. (Left) Coarse level. (Right) Fine level. (b) Horizontal clique message inheritance. (Up) Coarse level. (Down) Fine level.

VI. Experiments

Some practical experiments are carried out to test the AGBP method in this study. One of the typical applications is for stereo matching, which is a classical problem in computer vision. The purpose is to find a dense correspondence between two images that are captured in two different viewpoints. To simplify this process, the images are transformed to guarantee that two matched points only have a horizontal displacement, and then

TABLE I
EVALUATION OF EFFICIENCY AND CONVERGENCE

	Algorithm	Duration	Convergence energy	Acceleration rate
GBP	Canonical GBP	35482.5 s (≈10 hours)	406013	1.0 (Standard)
GBP+	+ caching	5198.8 s (≈1.5 hours)	406013	6.83 times
GBP++	+ caching + direction set	1969.815 s (≈30 min)	409994	18.01 times
AGBP	+ caching + direction set + hierarchical state space reduction	168.231 s (≈2 min)	381793	**210.92 times**

TABLE II
EVALUATION OF ACCURACY (%)

	e ∈[0, 1]	e ∈ (1, 2]	e ∈ (2, 3]	e > 3 pixels
Efficient BP	89.2	7.0	0.3	3.5
Canonical GBP	79.1	13.2	2.8	4.9
AGBP	86.5	8.1	1.8	3.6

the result also called depth map can be coded in a one-channel image. The gray value of a pixel represents the horizontal displacement of its matched pixel. There are a set of standardized image pairs to test stereo matching algorithms available on the web [15]. In the experiments on these image pairs, we use a set of parameters as: $T = 30.0$ and $\lambda = 0.87$ in the data cost term, $K = 10.0$ in the smooth cost term, and 5 for the number of levels from the coarsest to the finest. d and η in Γ depend on disparity range of specific image pair To make the data cost more smooth, the source image pairs are initially filtered by a 3×3 Gaussian smooth kernel with $\sigma = 0.7$.

To show the improved efficiency as well as the accuracy of the proposed method, the evaluation presented here is mainly among three algorithms, i.e., efficient BP [9], canonical GBP [11], and the proposed AGBP. The efficient BP, up to now, is the first and the only global algorithm that has achieved real-time demand in stereo matching using a parallel implementation [10]. Its efficiency and accuracy can be considered as a milestone of such message based algorithms. The implementation of our proposed method is by means of a single thread and therefore it is not reasonable to compare it with [10]. However, we believe our approach has some favorable features in the final results from a message-based point of view. The experiments of canonical GBP algorithm carried out in this research are aimed to show the typical accuracy of this kind of methods for comparison. In all of the experiments here, the same test set of image pairs provided by Scharstein and Szeliski [15] are used. The performance was tested by a commodity CPU with 2.13 GHz and 2G DRAM.

Among the three main optimization strategies, i.e., caching, direction set, and hierarchical state-space reduction, caching is a lossless method while direction set and hierarchical state-space reduction may cause loss of accuracy. The ultimate purpose of the proposed strategies is to improve the efficiency of canonical GBP as well as keep it in a good accuracy. As shown in Table I, the execution time which combines the three strategies can be extensively reduced, while the convergence energy rises a little bit due to some side effects. The canonical GBP augmented with all those three strategies can achieve about 200 times of the speed rate. It should be noted that the accelerating rates are related to the size of the MRFs and the range of the state space, not the image context [16]–[24]. The experiments were tested with "Tsukuba" (384×288 size) and "Venus" (434×383 size). The ranges of the state space (d in Γ) are 16 and 20, and the reduction constants (η in Γ) are 3 and 4.

Furthermore, since AGBP is an approach based on message passing, it can also be implemented in a parallel way. Attributing to the high efficiency of a parallel computing architecture, e.g., compute unified device architecture, it is found that the proposed method can execute much faster, even possibly for real-time applications.

Table II gives the accuracy evaluation of "Tsukuba" of three different algorithms. Since the proposed method uses a hierarchical strategy to encourage the information flowing, the accuracy is higher than that of canonical GBP. Comparing it with the accuracy of efficient BP, the proposed method yields a similar level. According to the results of [11], the acceleration rate is very small when the number of labels is below 64, i.e., from 0.6 to 5 times of the Standard speed. The rate goes up to 26 when the number of labels increases to 256. In stereo matching, however, the number of labels is usually from 10 to 30. By comparison, the proposed AGBP in this paper can always reach at about 200 times of acceleration.

Except for the errors of the results, another reasonable assessment criterion is to take a close look at the depth maps and analyze their visual quality. The comparison between the proposed method and canonical GBP is carefully assessed. Fig. 10 lists the results based on three test images which are provided by [15]. From the results, it is noticeable that each method has its own merits. If we concern some conspicuous feature points, the canonical GBP makes a better result than the proposed method. However, in some other areas, like the massive noise pervaded in the depth maps, since the proposed method uses a hierarchical structure to send information over a wider range, the surface becomes more accurate than that of canonical GBP.

On the other hand, through the comparison between the proposed method and efficient BP, it is noticeable that efficient BP tends to get a fronto-parallel result which makes the surface over smooth and results in a layered effect, although it can still achieve a good result on objects' boundary. On the contrary, the proposed method does not have the drawback of layered effects like that caused by efficient BP, but the depth map becomes blurred at the boundaries and some noises cannot be eliminated. In fact, although a layered result can reach a lower energy, it cannot always be a better description of the outdoor scenes. Since the test images are mostly composed of plane-like objects, the implied fronto-parallel assumption always gives the least cost in these situations. However, comparing with the proposed method, it causes some over sharp edges and obvious layered effect surely which are not realistic and visual favorite.

Table III gives the specific temporal analysis in iterations. The massive computation of canonical GBP is split into several

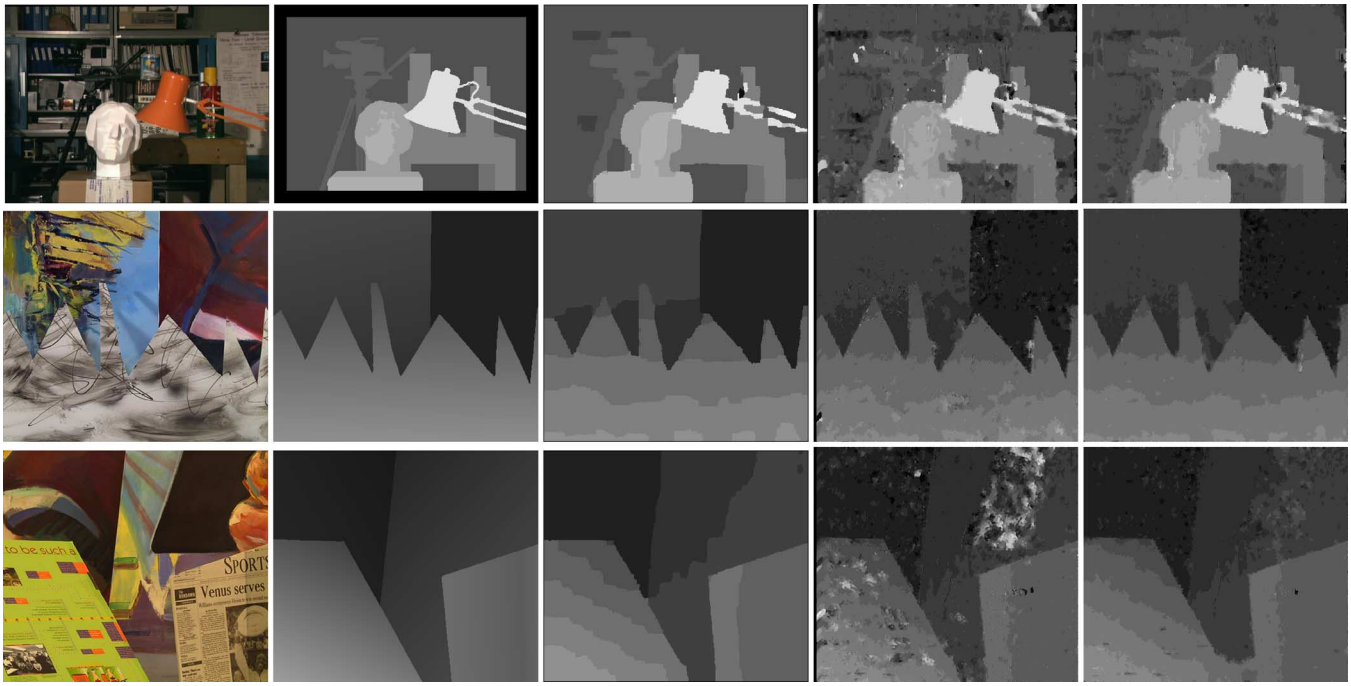


Fig. 10. Quality evaluation. From left to right: left images of the test pairs, ground truths, results of efficient BP, results of canonical GBP, and results of AGBP.

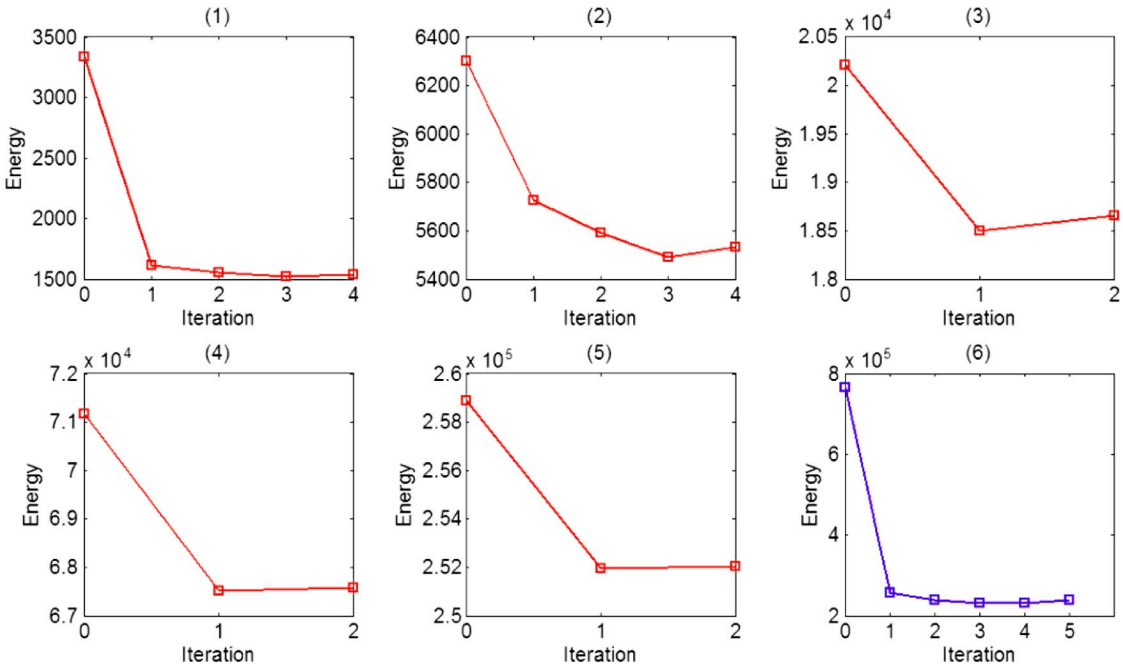


Fig. 11. Energy function in every iteration of the "Tsukuba" image pair. (1)–(5) Energy curves at each level. (6) Energy curve of canonical GBP.

small scale iterations in the proposed method. Since the canonical GBP carries the temporal complexity of $O(n^4)$, the duration of these small-scale iterations is sharply reduced. Taking a close look at the time costs at each level in the proposed method, the durations are still rising in times. The reason is that the parameter Γ is still very large at finer levels. However, if the labels are reduced more rapidly in Γ for the purpose of speed, the state space maybe shrinks too quickly so that the results could be worse. Therefore, the tradeoff between the speed and the accuracy is paradoxical. The settings used in our experiments have reached a good balance.

Fig. 11 shows the energy variation in canonical GBP and each level of the proposed method. From those curves, it can be found that the energy of canonical GBP decreases rapidly in the first iteration. In later iterations, the energy changes slowly. On the contrary, the energy in our proposed method decreases evenly in each iteration at each level, which can be treated as a space selection at different scales. The major effect of this selection

TABLE III
EXECUTION TIME IN ITERATIONS (ms)

Iterations	Proposed AGBP					Canonical GBP
	Level 5	Level 4	Level 3	Level 2	Level 1	
1	982	2808	5772	11076	17379	7299×10^3
2	889	2808	5756	11170	17706	7098×10^3
3	874	2761	5647	11029	17347	7177×10^3
4	858	2761	5632	10983	17097	6926×10^3
5				10998		6980×10^3

process is to get a better convergence, while canonical GBP often trends to trap in a local minimum near the initial position.

VII. CONCLUSION

The inference problem is very important in statistical physics as well as in many other fields. This paper especially solves such a problem in large-scale MRFs using AGBP. A min-sum scheme is invented for the message computing process in AGBP. Furthermore, two other strategies are proposed for improving the efficiency, i.e., the direction set and hierarchical state-space reduction. Using the direction set method, the complex of computing the cluster message can be reduced from $O(n^4)$ to $O(n^3)$, and using hierarchical state-space reduction, an extra acceleration rate of $n/(2k)$ can be achieved in theory, where n is the state space and k is the number of iterations. Therefore, when compared with canonical GBP, the proposed AGBP can improve the efficiency up to 200 times for a typical field size of 300×300 in the problem of stereo matching.

REFERENCES

[1] A. Sinha and S. Gupta, "A fast nonparametric noncausal MRF-based texture synthesis scheme using a novel FKDE algorithm," *IEEE Trans. Image Process.*, vol. 19, no. 3, pp. 561–572, Mar. 2010.

[2] D. Bruckner and R. Velik, "Behavior learning in dwelling environments with hidden Markov models," *IEEE Trans. Ind. Electron.*, vol. 57, no. 11, pp. 3653–3660, Nov. 2010.

[3] C. Wolf, "Document ink bleed-through removal with two hidden Markov random fields and a single observation field," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 3, pp. 431–447, Mar. 2010.

[4] S. S. H. Zaidi, S. Aviyente, M. Salman, S. Kwang-Kuen, and E. G. Strangas, "Prognosis of gear failures in DC starter motors using hidden Markov models," *IEEE Trans. Ind. Electron.*, vol. 58, no. 5, pp. 1695–1706, May 2011.

[5] L. Zhu, Y. Chen, and A. Yuille, "Unsupervised learning of probabilistic grammar-Markov models for object categories," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 114–128, Jan. 2009.

[6] J. Goldberger and H. Kfir, "Serial schedules for belief-propagation: Analysis of convergence time," *IEEE Trans. Inf. Theory*, vol. 54, no. 3, pp. 1316–1319, Mar. 2008.

[7] J. S. Yedidia, W. T. Freeman, and Y. Weiss, "Generalized belief propagation," *Neural Inf. Process. Syst.*, vol. 13, pp. 689–695, 2000.

[8] J. S. Yedidia, W. T. Freeman, and Y. Weiss, "Constructing free-energy approximations and generalized belief propagation algorithms," *IEEE Trans. Inf. Theory*, vol. 51, no. 7, pp. 2282–2312, Jul. 2005.

[9] P. Felzenszwalb and D. Huttenlocher, "Efficient belief propagation for early vision," *Int. J. Comput. Vis.*, vol. 70, no. 1, pp. 41–54, 2006.

[10] Q. Yang, L. Wang, R. Yang, H. Stewenius, and D. Nister, "Stereo matching with color-weighted correlation, hierarchical belief propagation, and occlusion handling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 3, pp. 492–504, Mar. 2009.

[11] K. Petersen, J. Fehr, and H. Burkhardt, "Fast generalized belief propagation for MAP estimation on 2D and 3D grid-like markov random fields," in *Proc. 30th Deutsche-Arbeitsgemeinschaft-fur-Mustererkennung Symp. Pattern Recognit.*, Munich, Germany, Jun. 2008, pp. 10–13.

[12] S. Y. Chen, H. Tong, Z. J. Wang, S. Liu, M. Li, and B. Zhang, "Improved generalized belief propagation for vision processing," *Mathematical Problems Eng.*, vol. 2011, 2011, art 416963, 12 pages.

[13] B. S. R. Armstrong and S. K. P. Veettil, "Soft synchronization: Synchronization for network-connected machine vision systems," *IEEE Trans. Ind. Informat.*, vol. 3, no. 4, pp. 263–274, Nov. 2007.

[14] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Reciples in C*. Cambridge, U.K.: Cambridge Univ. Press, 1992.

[15] D. Scharstein and R. Szeliski, Middlebury Stereo Vision Research Page [Online]. Available: http://vision.middlebury.edu/stereo/eval/ 2008

[16] Y. Sung, H. V. Poor, and H. Yu, "How much information can one get from a wireless ad hoc sensor network over a correlated random field?," *IEEE Trans. Inf. Theory*, vol. 55, no. 6, pp. 2827–2847, Jun. 2009.

[17] M. J. Choi, V. Chandrasekaran, and A. S. Willsky, "Gaussian multiresolution models: Exploiting sparse Markov and covariance structure," *IEEE Trans. Signal Process.*, vol. 58, no. 3, pp. 1012–1024, Mar. 2010.

[18] E. I. Athanasiadis, D. A. Cavouras, D. T. Glotsos, P. V. Georgiadis, I. K. Kalatzis, and G. C. Nikiforidis, "Segmentation of complementary DNA microarray images by wavelet-based markov random field model," *IEEE Trans. Inf. Technol. Biomed.*, vol. 13, no. 6, pp. 1068–1074, Nov. 2009.

[19] P. P. Gajjar and M. V. Joshi, "New learning based super-resolution: Use of DWT and IGMRF prior," *IEEE Trans. Image Process.*, vol. 19, no. 5, pp. 1201–1213, May 2010.

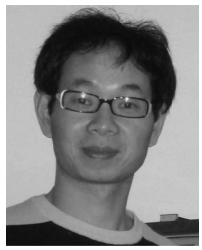
[20] C.-Y. Chung and H. H. Chen, "Video object extraction via MRF-based contour tracking," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 1, pp. 149–155, Jan. 2010.

[21] O. Dikmen and A. T. Cemgil, "Gamma Markov random fields for audio source modeling," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 3, pp. 589–601, Mar. 2010.

[22] J. Zhu, L. Wang, J. Gao, and R. Yang, "Spatial-temporal fusion for high accuracy depth maps using dynamic MRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 5, pp. 899–909, May 2010.

[23] M. Soccorsi, D. Gleich, and M. Datcu, "Huber-Markov model for complex SAR image restoration," *IEEE Geosci. Remote Sens. Lett.*, vol. 7, no. 1, pp. 63–67, Jan. 2010.

[24] I. Morgan and H. Liu, "Predicting future states with n-dimensional Markov chains for fault diagnosis," *IEEE Trans. Ind. Electron.*, vol. 56, no. 5, pp. 1774–1781, May 2009.



Shengyong Chen (M'01–SM'10) received the Ph.D. degree in computer vision from the City University of Hong Kong, Kowloon, Hong Kong, in 2003.

He joined the Zhejiang University of Technology, Hangzhou, China, in February 2004, where he is currently a Professor in the Department of Computer Science. From August 2006 to August 2007, he received a fellowship from the Alexander von Humboldt Foundation of Germany and worked at the University of Hamburg, Hamburg, Germany. From Sep. 2008 to Aug. 2009, he was a Visiting Professor at Imperial College, London, U.K. He is the author or coauthor of more than 100 scientific papers published in international journals and conferences. His research interests include computer vision, robotics, 3-D object modeling, and image analysis.

Dr. Chen is a Committee Member of IET Shanghai Branch.

Zhongjie Wang received the B.Sc. and M.Sc. degrees in pattern recognition and machine intelligence from the College of Information Engineering, Zhejiang University of Technology, Hangzhou, China, in 2007 and 2010, respectively, under the supervision of Dr. S. Chen.

He is currently a Visiting Scholar in Max-Planck-Institut für Informatik, Saarbruecken, Germany, under the supervision of Dr. T. Thormaelen. During his study, he joined a National Natural Science Foundation of China, "Purposive Perception Planning Method for 3D Target Modeling," completed his thesis on "Stereo Matching for 3D Computer Vision," and applied two patents.