# Novel 3D Video Conversion from Down-Sampled Stereo Video

Wun-Ting Lin
*Department of Computer Science*
*National Tsing Hua University*
*Hsinchu, Taiwan*
*Email: shiaushiauhan@gmail.com*

Shang-Hong Lai
*Department of Computer Science*
*National Tsing Hua University*
*Hsinchu, Taiwan*
*Email: lai@cs.nthu.edu.tw*

***Abstract*—Stereo video has become the main-stream 3D video format in recent years due to its simplicity in data representation and acquisition. Under stereo settings, the twin problems of video super-resolution and high-resolution disparity estimation are intertwined. In this paper, we present a novel 3D video conversion system that converts down-sampled stereo video to high-resolution stereo sequences with a Bayesian framework. In addition, we estimate the finer-resolution disparity maps with a two-step CRF model. Our super-resolution system can also be incorporated into the video coding process, which can significantly lower the data amount as well as preserving high-quality details. Experimental results demonstrate that our system can enhance image resolution in both stereo video and disparity map. Objective evaluation of the proposed video coding scheme combined with super-resolution at different compression ratios also shows competitive performance of proposed system for video compression.**

***Keywords*-Stereo video super-resolution; diparity refinement; video compression;**

## I. Introduction

3D video has become a popular trend in the entertainment industry in recent years, especially 3D movies. Advanced 3D display technologies allow users to experience realistic 3D effects at home. More and more 3D display applications can be found in the high-end electronic products, such as 3D LCD/LED TVs, 3D cameras, 3D camcorders, 3D laptops, 3D mobile phones, and 3D games, etc. The revolution from 2D display to 3D display has started to change many aspects of our daily life, including entertainment, communication, photography, and medical science. Many investigators and investors regard this frontier technology as great potential market in the near future. The service of providing massive 3D video content for people to watch will play an important role, and we need new technology to broadcast the 3D content as well as to display 3D videos with satisfactory quality so that users can enjoy the 3D services effortlessly.

On the other hand, multi-frame super resolution, namely estimating fine-resolution frames from a coarse-resolution sequence shown in Fig. 1, is one of the fundamental problems in computer vision. With the rapid advancement of display devices, such as HDTV or 4K2K-TV, super-resolution technique plays an extremely vital role nowadays, helping convert plenty of low-resolution multimedia content into high-resolution version. Moreover, following the popularization of mobile devices, super resolution is indispensable in building the bridge between display devices and digital cameras in mobile phones.

However, although super-resolution has been extensively studied for decades, applying super-resolution to real video sequences still remains quite challenging. Most previous works are sensitive to their assumed models of data and noise, which limit their approaches from practical application. In addition, for 3D stereo video, the sub-pixel registration information required for super-resolution is tightly coupled to the 3D structure, which also increases the complexity of the problem.

Therefore, a practical super-resolution system should take 3D structure into consideration and simultaneously estimate optical flow, noise level and blur kernel along with reconstructing the high-resolution frame. With the gradually maturing technique in each of these problems, it is natural to combine all these components into a single framework without making oversimplified assumptions.

In this paper, we present a novel 3D video conversion system that converts down-sampled stereo video into high-resolution stereo video sequence. In addition, we estimate the high-resolution disparity maps with finer details. The system first employs a stereo matching algorithm to compute the disparity for down-sampled stereo pairs. Then, a video super-resolution approach using the Bayesian framework is applied. The framework alternatively reconstructs the high-resolution frames as well as estimates the optical flow, noise level and the blurring kernel. Once the system estimates the fine-resolution stereo video, it combines the disparity maps computed in different resolution with proposed two-step conditional random field model, and generates a new disparity map which is more accurate in either depth value or object boundary.

Furthermore, we also propose to integrate our super-resolution system into the video coding process by encoding down-sampled sequence along with several high-resolution key-frames. Our system is exploited to achieve video coding at different compression ratios by adjusting the number of key-frames, and the objective analysis in PSNR and SSIM shows that this combination of super-resolution with video


Figure 1. Stereo video super resolution: the upscaled results on the right are simulated by the proposed system.

coding can efficiently lower the bitrate but still preserve delicate details in the video.

## II. Related Work

Video super-resolution has been extensively studied in the computer vision, image processing and computer graphics communities. The methods developed over the decades differ in their formulations, underlining prior models and the problem settings. The most representative and simplest approach is the interpolation-based methods, which attempt to predict intermediate unknown pixels with linear interpolation filter, such as the bilinear filter or bicubic filter. These interpolation kernels are designed for spatially smoothing which often conflicts with real-world image property that contains singularities, such as edges and high-frequency textured regions. Therefore, these interpolation-based methods suffer from various edge-related visual artifacts including ringing, aliasing, jaggy and blurry effects.

More sophisticated methods can be found in Park et al. [1], which provided a comprehensive technical survey of super-resolution techniques as well as their respective limitation and evaluation. Milanfar [5] introduced numerous approaches that have been successfully used in super-resolution in recent years. One of the main-stream approaches in multi-frames super resolution is the reconstruction-based method, which is based on the same concept employed in the proposed system. Tipping et al. [2] showed that estimating the motion between images with Bayesian approach instead of cross-validation can be more adaptive to the real-world images, which are under unknown point spread function blurring. Sun et al. [3] assumed that high-resolution images hold approximately the same total variance in gradient domain with low-resolution one, and reconstruct the result with modified gradient prior together with an external dataset. Liu et al. [4] proposed to concurrently estimate the optical flow, noise level and blur kernel in addition to reconstruct the target high-resolution frame, which suffers from the heavy computational load in calculating the dense flow between video frames. In spite of the satisfactory results by using these reconstruction-based methods, the registration information required in super resolution often yields undesirable artifacts due to errors in motion estimation.

Further, Bhavsar et al. [6] intended to enhance the resolution of stereo images in both color value and disparity with the multiple stereo inputs. They adopted the reconstruction-based idea and followed the conventional image capturing formation to formulate the stereo-image super-resolution problem. Although their disparity results are quite fascinating, their high-resolution color images suffer from image blurring and unsatisfactory artifacts for the sake of oversimplified model in noise level and blur kernel. Zhang et al. [7] proposed a closed-loop super-resolution method for multi-view stereo consisting of numerous low-resolution images and a single high-resolution image. Their method first employs a stereo matching technique and fuses the multiple disparities into a unique depth map. Then, a super-resolve approach that predicts the target-view image under the guidance of the depth information is presented. Nonetheless their results are satisfactory, the necessary high-resolution input image is uneasy to obtain in practice, thus limiting this approach from real-world application.

## III. Stereo Video Super-Resolution with Bayesian Framework

Given a low-resolution stereo sequence $\{J_t^L,\ J_t^R\}$, captured from two-fixed view named left and right, our goal is to predict the high-resolution stereo video $\{I_t^L,\ I_t^R\}$. For simplicity, we take only the left view as our modeling object despite the other view can also be formulated in the similar manner without excessive modification. In consideration of computational complexity, our approach estimates the high-resolution frame $I_t^L$ only with the limited adjacency frames

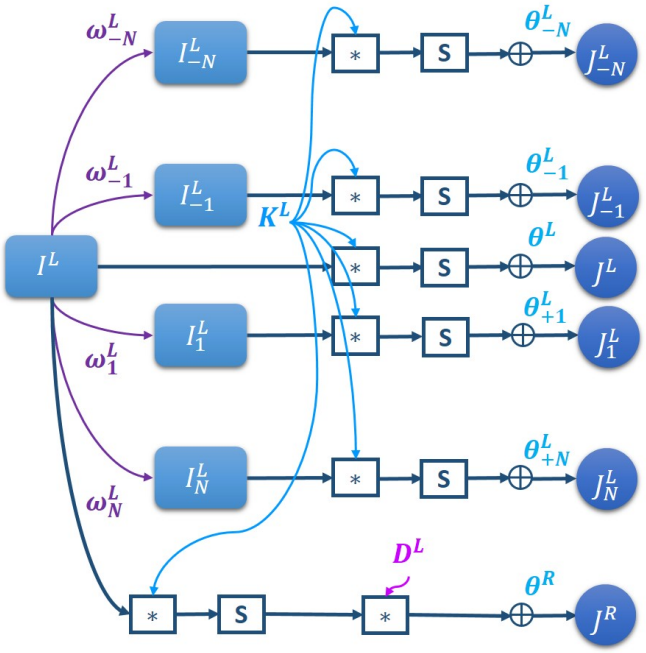


Figure 2. This is the graphical model of our stereo video super resolution problem. The observed coarse resolution images in left view go through motion, blur and noise processes to become $J^L$ while $J^R$ is passing one-more warped process by the corresponding disparity value.

$\{J^L_{t-N}, \ldots, J^L_t, \ldots, J^L_{t+N}\}$ and a single opposite-view frame $J^R_t$ located at time t. Again, we omit the subscript $t$ from now for easy of exposition. As mentioned in section I, in order to better approximate the image formation process, our approach not only estimates the desired high-resolution video sequence but also models the motion, noise level and blur kernel.

We represent the motion between frames belonging to same view with optical flow $\{\omega^L_i\}$, which indicates the dense flow from target frame to the $i^{th}$ adjacency frame, holding the vital registration information required in super-resolution. Besides, noise level in our framework is depicted as an uncertain variable owing to the outlier generated in flow estimation, which was also estimated with Bayesian approach and written as $\{\theta^L_i, \theta^R\}$ for two different sources of reference frame, respectively. Without loss of generality, our system also estimates the blur kernel $K^L$ as well, which is related to the point spread function in the camera capture process. The graphical model illustrating the stereo video super-resolution problem can be found in Fig. 2, where $S$ stands for the down-sampled operator that uses the average operator, and $D^L$ is the estimated disparity map.

## IV. Stereo Video Super-Resolution Video Super-Resolution with Bayesain Framework

First, our approach employs a stereo matching method to calculate the disparity $D^L$ from left view to right view. To compromise the tradeoff between performance and execution time, our system applies Yang's [8] stereo matching method to generate the required disparity map for low-resolution stereo pair at time $t$.

After obtaining an initial disparity map, the stereo video super-resolution problem can formulated in the Bayesian MAP framework as follows:

$$\left\{I^{L^*}, \left\{\omega^{L^*}_i\right\}, K^{L^*}, \left\{\theta^{L^*}_i\right\}, \theta^{R^*}\right\} = \arg\max p\left(I^L, \{\omega^L_i\}, K^L, \{\theta^L_i\}, \theta^R \middle| \{J^L_i\}, J^R, D^L\right) \tag{1}$$

By the well-known Bayesian theorem, the posterior probability in eq. (1) can be decomposed into a series of multiplication consists of likelihood and prior, which is given in eq. (2).

$$\begin{aligned} p\left(I^L, \{\omega^L_i\}, K^L, \{\theta^L_i\}, \theta^R \middle| \{J^L_i\}, J^R, D^L\right) \propto \\ p\left(I^L\right) p\left(K^L\right) p(\theta^R) \prod_i p\left(\omega^L_i\right) \prod_i p\left(\theta^L_i\right) \\ \cdot p\left(J^L \mid I^L, K^L, \theta^L\right) p(J^R | I^L, K^L, D^L, \theta^R) \\ \cdot \prod_{i \neq 0} p\left(J^L_i \mid I^L, K^L, \omega^L_i, \theta^L_i\right) \end{aligned} \tag{2}$$

Motivated by the success of Liu et al. [4] in single-view video super-resolution, we solve our problem in the similar manner, which divides the problem in eq. (2) into several sub-problems and solves each of them alternatively with the IRLS optimization technique.

### A. High-resolution image estimation

Here we are going to show that with the initial estimation of optical flow, noise level and blur kernel, eq. (2) can be simplified to have only one unknown variable, the high-resolution image. Some previous works [9][10] have demonstrated the advantageous of sparse prior in preserving the singularities, such as edges or high-frequency textured regions, in natural images. Therefore, our system also adopts the sparse prior to regularize our output images. Objective function for high-resolution image can be found in eq. (3), which follows the formulation in the afore-mentioned graphical model in Fig. 2, where S, $F_{flow}$, $K$ stand for the matrix of down-sample process, motions from either optical flow or different viewpoint and the matrix for convolution with a point spread function , respectively.

$$\begin{aligned} I^{L^*} = \arg\min \theta^L \left\|SK^L I^L - J^L\right\| + \eta \left\|\nabla I^L\right\| \\ + \theta^R \left\|F_{D^L} SK^L I^L - J^R\right\| + \sum_{i \neq 0} \theta^L_i \left\|SK^L F_{\omega^L_i} I^L - J^L_i\right\| \end{aligned} \tag{3}$$

To solve eq. (3), we rewrite the equation to contain only a single term in the objective function and approximate the

Manhattan norm with IRLS, which is shown in eq. (4). The symbol $G_\nabla$ represents the derivative operator.

$$I^{L*} = \arg\min \begin{bmatrix} \theta^L \\ \theta^R \\ \eta \\ \theta_i^L \end{bmatrix}^T \left\| \begin{bmatrix} J^L \\ J^R \\ \overrightarrow{0} \\ J_i^L \end{bmatrix} - \begin{bmatrix} SK^L \\ F_{D^L} SK^L \\ G_\nabla \\ SK^L F_{\omega_i^L} \end{bmatrix} I^L \right\|, \quad \text{for i} \in [-\text{N},\ \text{N}],\ \text{i} \neq 0 \tag{4}$$

In the process to find the estimate of $I^L$, our system automatically eliminates the flow information if it is not adequately reliable. Namely, we evaluate the reliability of flow $\omega_i(x)$ in each position $x$, where $\omega_i$ represents the motion from target frame to reference frame. First, we generate the inverse flow $\widetilde{\omega}_i(x)$, which describes the motion from reference frame back to target frame. Next, a pixel $x$ in target frame is warped to the corresponding position $x$' in reference frame with the flow $\omega_i(x)$ information. Then, we warp $x'$ back to target frame with previous calculated inverse flow$\widetilde{\omega}_i(x')$ to get $x''$. A simple threshold examination is utilized here to inspect the reliability of flow $\omega_i(x)$ and the threshold value is adjusted by the upscaling factor in super-resolution, which is usually set to 2. In short, in eq. (4) we only take the reliable flow information to generate our desired images.

### B. Noise level estimation

In our system, we take the Gamma distribution in eq. (5) as the prior for the noise level,

$$\text{p}\left(\theta_i; \alpha,\ \beta\right) = \frac{\beta^\alpha}{\Gamma(\alpha)} \theta_i^{\alpha-1} e^{-\theta_i \beta} \tag{5}$$

where $\theta_i$ has a close form solution derived by Liu et al. [4] under the determined estimation of high-resolution images, optical flow and blur kernel. Equation (6) and (7) show the solution for two different sources of reference frames, $J_i^L$ and $J_0^R$, respectively.

$$\begin{aligned} \theta_i^{L*} &= \frac{\alpha + N_q - 1}{\beta + N_q \overline{x}^L} \\ ,\overline{x}^L &= \frac{1}{N_q} \sum_{q=1}^{N_q} \left| \left( SK^L F_{\omega_i^L} I^L - J_i^L \right)(q) \right| \end{aligned} \tag{6}$$

$$\begin{aligned} \theta^{R*} &= \frac{\alpha + N_q - 1}{\beta + N_q \overline{x}^R} \\ ,\overline{x}^R &= \frac{1}{N_q} \sum_{q=1}^{N_q} \left| \left( F_{D^L} SK^L I^L - J^R \right)(q) \right| \end{aligned} \tag{7}$$

In the above two equations, symbol $q$ represents pixel index while $N_q$ stands for the total number of pixels.

### C. Motion estimation

To avoid estimating the optical flow across two different resolutions, we scale up the reference frames with the inverse matrix of $S$ and $K^L$ to get $\tilde{j}_i^L$ that has the same resolution as that of high-resolution image $I^L$. Here, we apply Liu's method [11] to estimate the flow between $\widetilde{J_i^L}$ and $I^L$ for our super-resolution framework.

### D. Blur kernel estimation

We assume that blur kernel $K^L$ can be separated into two one-dimensional filters $K_x^L$ and $K_y^L$, along with x-directional and y-directional, respectively. The problem of finding the blur kernel $K^L$ can be replaced by solving the problem of estimating $K_x^L$ and $K_y^L$ instead. Once more, accompanying with the fixed estimation of high-resolution images $I^L$, low-resolution frame $J^L$ and its noise level $\theta^L$, we derivate the objective function for estimating the blur kernel by incorporating the sparse prior constraint in eq. (8), which can be solved with a similar way for eq. (4). On this spot, we only show the equation for x-direction blur kernel for representative. In eq. (8), $A^L$ is the matrix composited by $I^L$, serving as the convolution operator such that $I^L \bigotimes K^L = A^L K^L$, and $M_y^L$ stands for the convolution operator between $K_x^L$ and $K_y^L$, which means $M_y^L K_x^L = K_x^L \bigotimes K_y^L$.

$$K_x^{L*} = \arg\min \theta_0^L \left\| A^L M_y^L K_x^L - J_0^L \right\| + \xi \left\| \nabla K_x^L \right\| \tag{8}$$

The initial value of noise level is computed by the difference of respective low-resolution pairs while the high-resolution optical flow is initially determined with upscaled version of video frames, which uses the linear interpolation filter. Further, the original blur kernel is assumed to be standard Gaussain distribution. Thus, we deal with these sub-problems iteratively, which has only four hyper parameters $\eta,\ \alpha,\ \beta\ and\ \xi$.

## V. Disparity Refinement with Two-Step CRF Model

Besides enhancing the resolution of color image, our framework also integrates the method for improving the depth quality as well. Our system partitions the super-resolution processing into finer steps in order to collect more estimated images across different resolutions. For example, we divide the procedure of four-time enlargement into twice two-time upscaling tasks to estimate the middle-resolution images. Via this, for one stereo color image pair, we have three different disparity maps scattering from three resolutions: coarse, middle and fine, which are labeled as level three, two and one, respectively.

The strategy of disparity refinement combines two-step CRF models where the former modifies the disparity values with tree-structured CRF, using hierarchical information

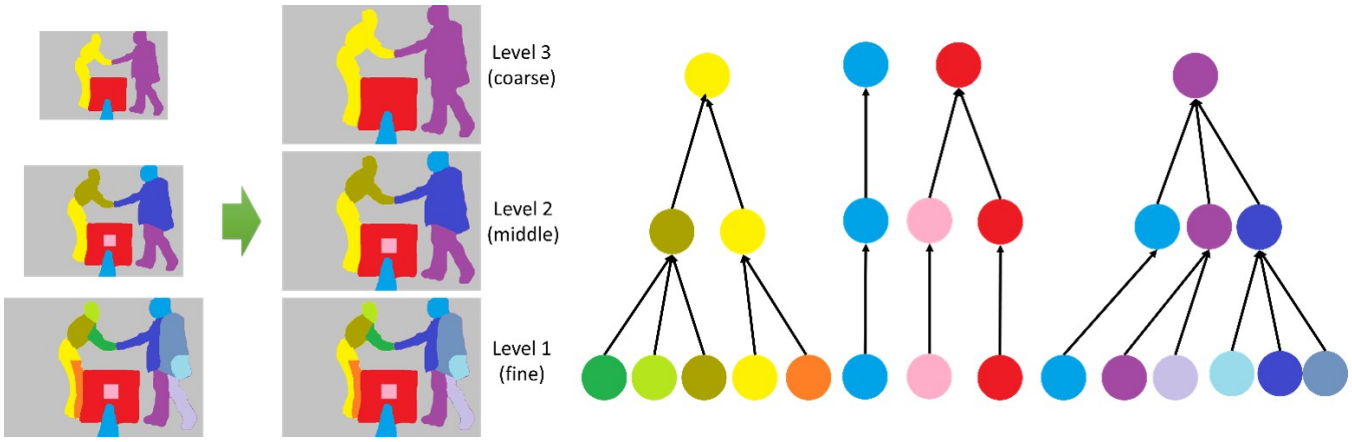


Figure 3. Tree-structured conditional random field. For each node which is a segment in level $l$, we build the tree by connecting the edge to only single node in level $l+1$ which has maximal overlapping pixels. With such simple tree-structured, the disparity refinement problem can be solved by message passing.

cross levels while the latter adjusts the object boundary with a fully connected CRF based on the results of step-one.

### A. Step one: Tree-structured conditional random field

Motivated by the success of Reynolds et al. [12] in figure-ground segmentation, we first attempt to enhance the disparity maps estimated from Yang [8] with tree-structured conditional random field, which are depicted in Fig. 3. Given three disparities scattered in earlier-stated three different levels, our goal is to generate a refined single disparity map by combining all information of neighbors and levels.

We begin with upscaling each disparity map to the highest resolution, the resolution of desired high-resolution images, and apply the over-segmentation technique by Liu et al. [13] to each disparity map to obtain segments, which correspond to nodes in our graphical model in step one. The $i^{th}$ segment in level $l$ is displayed as $s_i^l$ and the mean color of all pixels in $\mathrm{s}_i^l$ is depicted as $m_i^l$. For better reflecting the information in different disparities, we assign distinct segment numbers for different resolutions, which also follow fine-to-coarse manner for level 1, 2 and 3. Then, we build a tree with node $i$ at level $l$ connecting to a single node j at level $l+1$, where j denote the node for the region with maximal pixel overlap with $i$, described in eq. (9).

$$\mathrm{j}^* = \arg\max \frac{\left|V_i^l \cap V_j^{l+1}\right|}{\left|V_i^l\right|} \tag{9}$$

Subsequently, we construct a forest composed of trees whose amount is the number of segments in the coarsest level.

To become a standard CRF model, we discretize the value of our disparity maps into 80 labels, written as $\{\mathrm{y}_1, \mathrm{y}_2, \ldots, \mathrm{y}_{80}\}$, estimating the maximal possibility by message passing in eq. (10).

$$\mathrm{p}\left(\mathrm{y} \mid m_i^l\right) = \frac{1}{Z(m_i^l)} \prod_{<ij>} \psi_{ij}(y_i, y_j) \prod_i \phi_i(y_i) \tag{10}$$

As general, we define node potential in eq. (11) with truncated $\mathrm{L}_1$-distance, which relates to the distance between labels and $m_i^l$ and the edge potential in eq. (12), where $\lambda_{ij} = e^{-\chi_{ij}^2}$ and $\chi_{ij}$ measures the similarity between segments using Bhattacharyya distance. In eq. (12), $\lambda_{ij} \approx 1$ in the similar segments, while $\lambda_{ij} \approx 0$ in the dissimilar pairs.

$$\phi_i\left(y_i\right) = \mathrm{e}^{-\delta \min(\left|m_i^l - y_i\right|,\ 10)} \tag{11}$$

$$\psi_{ij} = \begin{cases} e^{\lambda_{ij} \cdot \gamma}, & if\ i = j \\ e^{-\lambda_{ij} \cdot \gamma}, & otherwise \end{cases} \tag{12}$$

Finally, we solve eq. (10) with belief propagation with the public software [16], which is fast and exact inference due to the simple tree structure of the proposed method, and we assigns the segment in finest disparity level with the label of maximal possibility, which comes out to be the initial map of the next step.

### B. Step two: Fully connected conditional random field

In multi-class image segmentation problem, Krhenbhl et al. [14] shows that the fully connected CRF accomplishes to model the object structure better than grid CRF in most cases, and they proposed a highly efficient inference algorithm for fully connected CRF. Motivated by the admirable results they achieved in segmentation, we resolve our problem that refines the disparity boundary of the preceding step in a similar manner by quantizing our disparity maps to 40 levels, and we can obtain an inference result by using the algorithm by Krhenbhl et al. [14] in constant time.

## VI. Super-Resolution Involved Coding Scheme

In this paper, we also demonstrate the efficiency to integrate super-resolution framework into coding processing. We achieve this by encoding the down-sampled stereo video with several original-resolution key-frame $\{I^{Key}\}$ whose amount controls the ratio of compression. With the addition information $I^{Key}$, the Bayesian MAP formulation can be modified to eq. (13).

$$\begin{aligned}\left\{I^{L^*}, \Omega^*, K^{L^*}, \Theta^*\right\} = \\ \arg\max p\left(I^L, \Omega, K^L, \Theta \middle| \left\{J_i^L\right\}, J^R, D^L, I^{Key}\right) \\ , \Omega = \left\{\omega^{Key}, \left\{\omega_i^L\right\}\right\}, \Theta = \left\{\theta^{Key}, \left\{\theta_i^L\right\}, \theta^R\right\}\end{aligned} \tag{13}$$

which can be modified into eq. (14) with Bayesian theorem.

$$\begin{aligned}p\left(I^L, \Omega, K^L, \Theta \middle| \left\{J_i^L\right\}, J^R, I^{Key}\right) \\ \propto p\left(I^L\right) p\left(K^L\right) \prod_{\omega\in\Omega} p\left(\omega\right) \prod_{\theta\in\Theta} p\left(\theta\right) \\ p\left(J^L \mid I^L, K^L, \theta^L\right) p\left(J^R \mid I^L, K^L, D^L, \theta^R\right) \\ p\left(I^{Key} \mid I^L, \omega^{Key}, \theta^{Key}\right) \prod_{i\neq 0} p(J_i^L | I^L, K^L, \omega_i^L, \theta_i^L) \\ , \Omega = \left\{\omega^{Key}, \left\{\omega_i^L\right\}\right\}, \Theta = \left\{\theta^{Key}, \left\{\theta_i^L\right\}, \theta^R\right\}\end{aligned} \tag{14}$$

We solve eq. (14) with the similar scheme used in section ‡V by dividing this problem into several sub-problems and we only show the final objective function here. First, the equation dealing with high-resolution images can be changed into eq. (15), which takes the message of key-frame into consideration.

$$\begin{aligned}I^{L^*} = \arg\min \theta^L \left\|SK^L I^L - J^L\right\| + \theta^R \left\|F_{D^L} SK^L I^L - J^R\right\| \\ + \theta^{Key} \left\|F_{\omega^{Key}} I^L - I^{Key}\right\| + \eta \left\|\nabla I^L\right\| \\ + \sum_{i\neq 0} \theta_i^L \left\|SK^L F_{\omega_i^L} I^L - J_i^L\right\|\end{aligned} \tag{15}$$

The noise level of key-frame can be calculated with close form solution again in eq. (16) under the Gamma distribution prior of $\theta^{Key}$.

$$\begin{aligned}\theta^{Key^*} = \frac{\alpha + N_q - 1}{\beta + N_q \overline{x}^{Key}} \\ , \overline{x}^{Key} = \frac{1}{N_q} \sum_{q=1}^{N_q} \left|(F_{\omega^{Key}} I^L - I^{Key})(q)\right|\end{aligned} \tag{16}$$

The high-resolution optical flow can be computed directly by estimating the fine-resolution image pair $I^L$ and key-frame $I^{Key}$ without applying linear interpolation operator for upscaling, which often lower the accuracy of optical flow. Therefore, with slight adjustment, our super-resolution can be incorporated into the video compression process, which is efficient in lowering the transmission bitrate together with improvement in video quality.

## VII. Experimental Results

We conduct several experiments to evaluate the efficiency of the proposed method described in section III, IV and V, respectively.

### A. Stereo video super-resolution

We use the 3D video dataset Bookarrival downloaded from the website http://sp.cs.tut.fi/mobile3dtv/stereo-video/ for experiments. In average, upscaling a single frame for four times from size (256,192) to (1024, 968) with four forward adjacency frames and four backward adjacency frames takes about half hour on the computer equipped with i5-2500 CPU @ 3.30GHz and 16.0GB memory. Moreover, we empirically set our parameter $\eta$, $\alpha$, $\beta$ and $\xi$ mentioned in section ‡V to be 0.01, 0.1, 0.1 and 0.01, respectively, across all the experiments.

Because most super-resolution approaches did not release executable code for reproducing their results, we only compare to those methods with computer programs for fair comparison on our own stereo datasets. We choose the method of Shan et al. [15], which is one of the most competitive approaches recently as well as the classical bicubic linear interpolation approach.

Fig. 4 shows the result of two selected frames in Bookarrival sequence which totally has 100 frames, and the average PSNR and SSIM values of all sequence are placed in Table I. Without any doubt, the proposed system can generate better results than other approaches, which support by the highest values in both PSNR and SSIM. For better comparison, the zoom-in version of Fig. 4 can be found in Fig. 5. Limited by the property of interpolation, bicubic interpolation tends to provide smooth results which reflect to their SSIM scores. On the other hand, although the better contrast reached by Shan et al. [15], the over-emphasized edge make it less consistent to the ground-truth appearance, such as the man's shirt in the first column of Fig. 5 is obviously over-enhanced. Furthermore, their results sometimes are accompanied with several undesired artifact, which can clearly be seen in the outline of the sticker written "FFI" appeared in the second column of Fig. 5.

### B. Disparity refinement

As mentioned in section IV, we take the super-resolution task in progressive manner by dividing the four-time enlargement into twice double-size upscaling tasks to include the middle level estimation. In Fig. 6, the above two columns are the disparity generated by Yang [8] from different resolution stereo pairs in two selected frames: fine, middle and coarse from left to right sequentially. The disadvantage of low-resolution disparity is the coarse and imprecise boundary

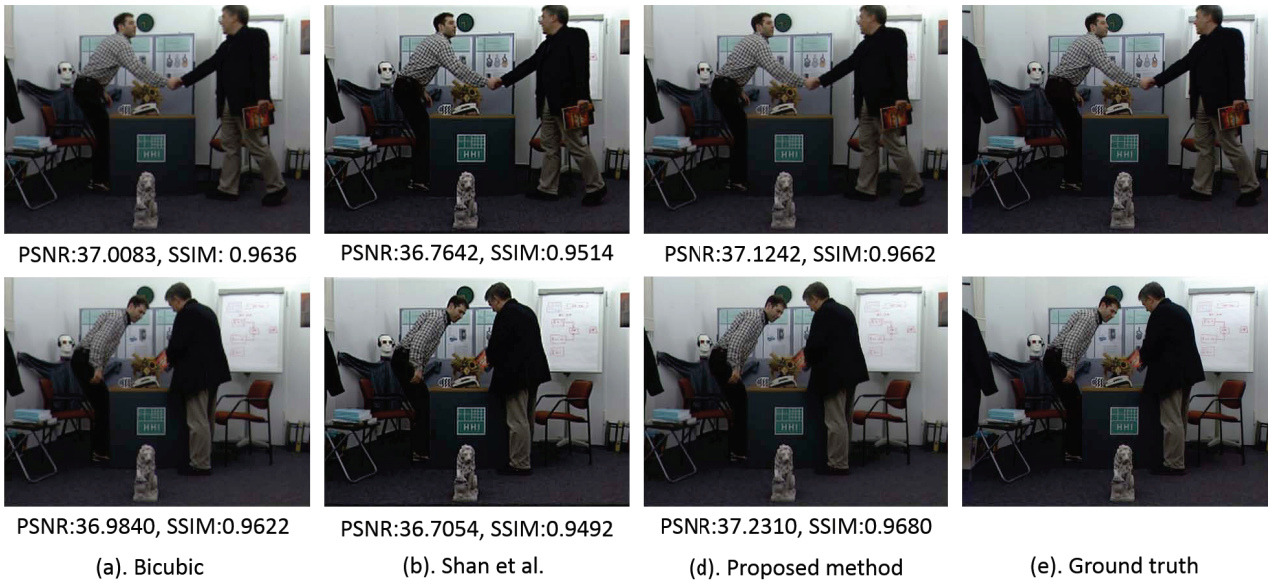


Figure 4. This is the comparison of our video super-resolution framework in four times enlargement with Bicubic and Shan et al. [15] Referring to Ground truth, the results generated by Bicubic tends to be blurring while the contrast in the estimated images of Shan et al. is so heavy that is slightly inconsistent to the original images. This compared figure is better seen in screen.

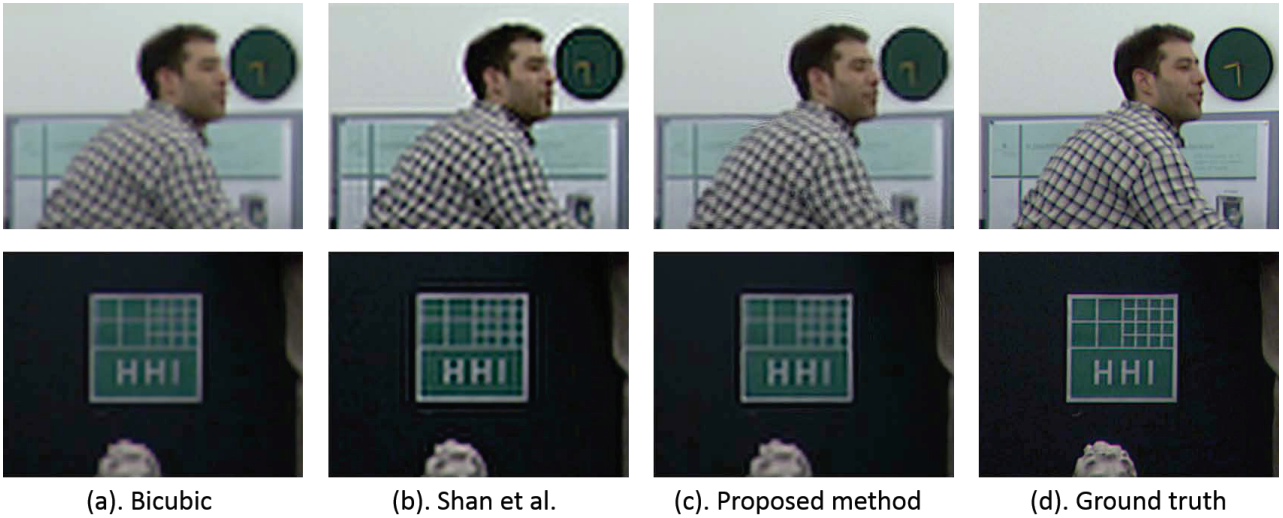


Figure 5. This is the zoom in version of Fig. 4, where we can find that the results of Shan et al. [15] suffer from the ringing artifacts, such as the boundary of the man in first column and the outline of the green sticker in the second column.

of object, while the drawback in high-resolution disparity is their increasing matching error due to the limitation of stereo-matching method.

The outcome of our first step in two-step disparity refinement is demonstrated in Fig. 6 (d), which strives to correct the value referring to all the disparities in levels with the tree-structured CRF. We state several apparent errors in the initial disparity map in Fig. 6 (a), (b) and (c), computed by Yang with red rectangle. After applying the step one refinement algorithm, the disparity values become more faithful that those red rectangle regions are all corrected. Then, we put the results in (d) to the fully connected conditional random field model introduced in section IV for adjusting the boundary of subject and then the results depicted in Fig. 6(e), which are successful in describing the outline of an object. Fig. 6 (f) depicts the corresponding color images computed by our two-step CRF model in measuring the similarity of either segments or pixels.

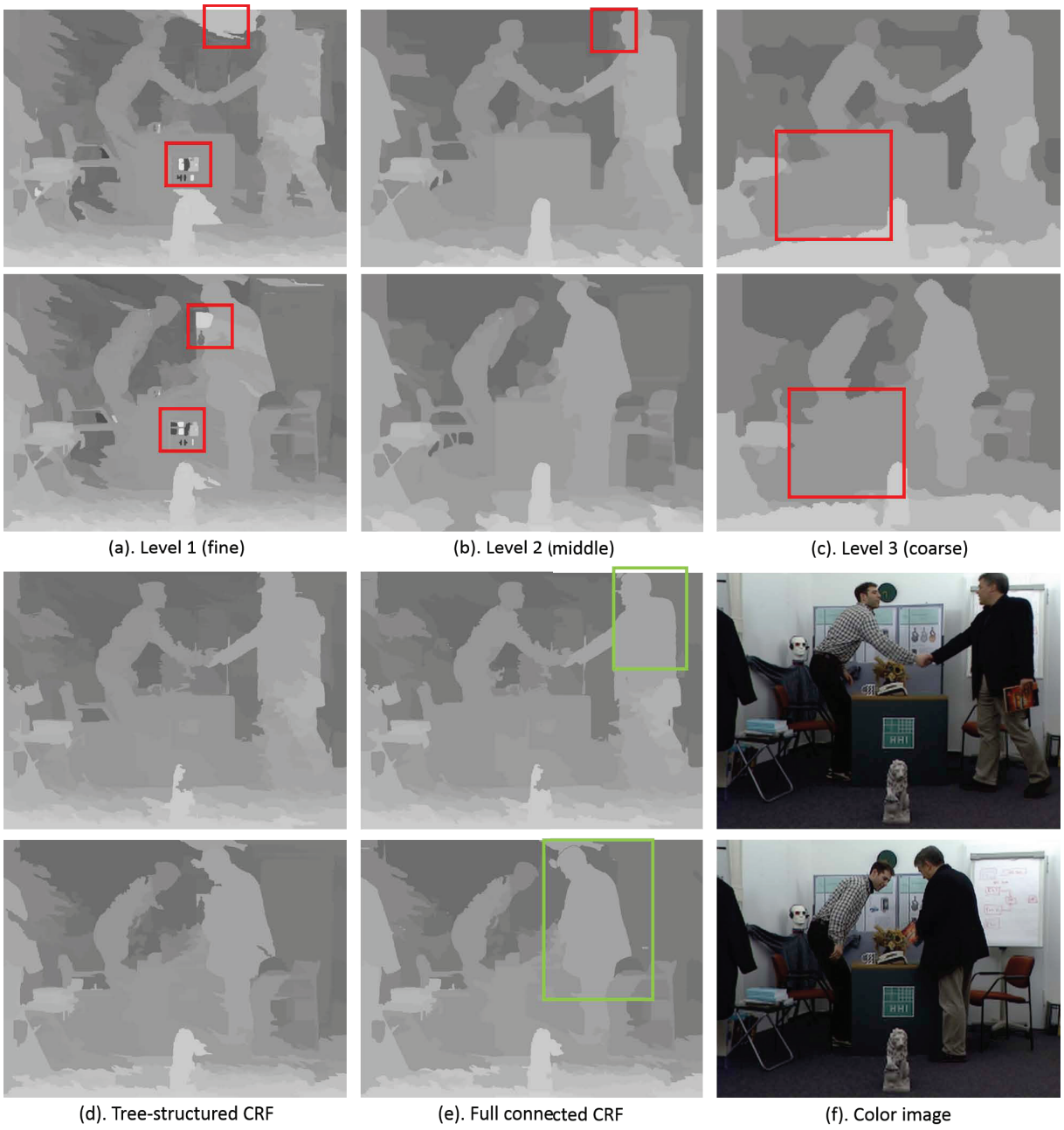


Figure 6. We shows the disparity refinement process of the two selected frames same in Fig. 4. From (a) to (c) are the disparities in level 1, 2 and 3 which are from coarse, middle to fine resolution respectively. The results (d) shows the step one estimation, namely tree-structured CRF, which targets for correcting the disparity error where stating as red rectangle. Our step 2 results of full connected CRF shown in (e), attempting to refine the boundary of objects stating with green rectangle. Then, (f) shows the corresponding color images used in both CRF model for measuring the similarity between either segments or pixels.

### *C. Results for super-resolution involved coding process*

Finally, we test our super-resolution involved coding system with different amount of key-frame for distinct compression ratio. Our coding process encodes two sequences, one is down-sampled video, which we use either factor 0.5 or factor 0.25 for experiments, while the other is the video containing original resolution frames whose total frame numbers ranging from 0 to 50 in step 5 for our 100 frames dataset. We use MATLAB built-in MPEG-4 compression method with frame rate 30 fps and quality 100 to compress both sequences. Fig. 7 shows the objective score of the super-resolution involved coding process with PSNR and SSIM.

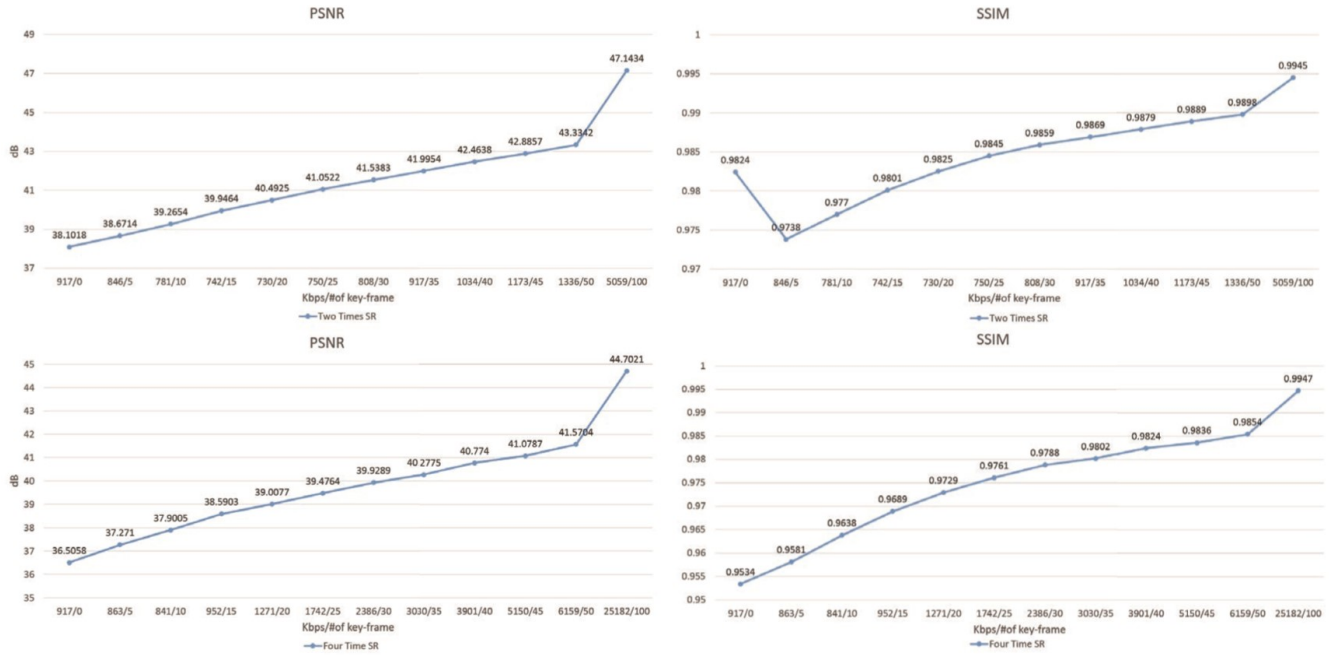


Figure 7. We test our super-resolution involved coding system with two different scale, 2X and 4X whose PSNR and SSIM value is shown here. Above figure represents the PSNR and SSIM value of scale two while the below represents the score of scale four. Both experiments prove that proposed method can lower lots bitrate but preserve certain quality values.

The horizontal direction indicates the different numbers of key-frames and the corresponding bit-rates after MPEG-4 compression, while the vertical shows the analysis values.

In fig. 7, the rightest point means the original classical MPEG-4 compression method, which has been shown to be quite successful in video coding. However, the proposed coding scheme can lower quarter bit-rate in scale of 0.5 and 0.25 with the PSNR value as high as 43.3342 and 41.5704 respectively.

Furthermore, in the scale 0.5 down-sampled stereo video, for the extremely case which contains 20 key-frames, the proposed method can lower the bit-rate up to 6.8 times with the PSNR value 40.4925. On the other hands, for the scale 0.25 down-sampled video, the extremely case which contains 10 key-frame can lower the bit-rate up to 30 times with the PSNR value 37.9005. This experiment shows that the proposed super-resolution involved coding scheme can dramatically lower the bitrate up to 30 times without losing too much image quality.

## VIII. Conclusion

In this paper, we proposed a Bayesian framework to solve the stereo video super-resolution problem together with estimating the optical flow, noise level and blur kernel for improving the stereo video quality and make it suitable for real-world sequences. After obtaining high-resolution stereo video, we apply our two-step CRF model to refine a disparity which can better describe the 3D scene in both depth value and object boundary. Moreover, the most important contribution this paper achieved is that we show that the super-resolution involved coding process together with standard MPEG-4 compression approach can dramatically minimize the transmission bitrate without losing too much image quality.

Although the proposed method can provide much better compression ratio, our system has lots of zoom for improvement in order to make it into practical use. Our future work will focus on the speed-up aspect by using the popular GPGPU platform as well as the optimization of our system to be more reliable from noise and outlier generated in optical flow and stereo matching.

## References

[1] S. C. Park, M. K. Park, and M. G. Kang, "Super-resolution image reconstruction: a technical overview," in *IEEE Signal Proc. Magazine,* 20(3):21–36, 2003.

[2] M. E. Tipping and C. M. Bishop, "Bayesian image super-resolution," "in *Advances in Neural Information Processing Systems,* 2002.

[3] J. Sun, Z. Xu, and H. Y. Shum, "Image super-resolution using gradient profile prior," "in *Computer Vision and Pattern Recognition,* 2008.

[4] C. Liu, D. Sun, "A Bayesian approach to adaptive video super resolution," in *Computer Vision and Pattern Recognition,* 2011.

[5] P. Milanfar, "Super-Resolution Imaging," in *Taylor I& Francis, CRC Press,* 2010.

[6] A. V. Bhavsar and A. N. Rajagopalan, "Resolution Enhancement in Multi-Image Stereo," in *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 32 (2010) 1721–1728.

[7] J. Zhang, Y. Cao, Z. Zheng and Z. Wang, "A New Closed Loop Method of Super-Resolution for Multi-view Images," in *International Conference on Multimedia Modeling,* 2013.

[8] Q. Yang, "A Non-Local Cost Aggregation Method for Stereo Matching," in *Computer Vision and Pattern Recognition,* 2012.

[9] M. Elad, M. A. T. Figueiredo and Y. Ma, "On the Role of Sparse and Redundant Representations in Image Processing," in *IEEE Special Issue on Applications of Compressive Sensing I& Sparse Representation,* 2010.

[10] J. Yang, J. Wright, Y. Ma and T. Huang, "Image super-resolution as sparse representation of raw image patches," in *Computer Vision and Pattern Recognition,* 2008

[11] C. Liu, "Beyond Pixels: Exploring New Representations and Applications for Motion Analysis," in *Doctoral Thesis in Massachusetts Institute of Technology,* 2009.

[12] J. Reynolds and K. Murphy, "Figure-ground segmentation using a hierarchical conditional random field," in *Fourth Canadian Conference on Computer and Robot Vision,* 2007.

[13] M. Y. Liu, O. Tuzel, S. Ramalingam and R. Chellappa, "Entropy Rate Superpixel Segmentation," in *Computer Vision and Pattern Recognition,* 2011.

[14] P. Krhenbhl and V. Koltun, "Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials," in *Neural Information Processing Systems,* 2011.

[15] Q. Shan, Z. Li, J. Jia, and C. K. Tang, "Fast image/video upsampling," in *ACM Transactions on Graphics (TOG),* 27(5):153, 2008.

[16] M. Schmidt, "UGM: Matlab code for undirected graphical models," at *http://www.di.ens.fr/mschmidt/Software/UGM.html,* July, 2013.