camera parameters with the rectification homography, which results in a so-called quasi-Euclidean rectification. The main difference between their algorithms and the proposed algorithm is that we improve the parameterization with a local optimization to avoid ambiguous solutions and undesirable rectification distortion. Moreover, the robust estimator is employed into the optimization process to overcome the inevitable outlier problem in the feature point correspondence. Lastly, we impose the temporal constraints in the updated image rectification for stereo video sequences.

Our algorithm aims to calibrate cameras as well as rectify images for uncalibrated stereo video sequences with temporally varying camera motions and zooming in/out effects. For the first frame, we estimate a reduced set of camera parameters through a nonlinear optimization process to minimize the geometric errors of the match points in pre-rectified image coordinates. For the subsequent frames, we formulate an objective function that jointly considers the geometric errors and smoothness constraints over temporal variations, and the camera parameters are updated so as to minimize the objective function. In this framework, the proposed algorithm contains the following advantages: 1) temporal stability while varying camera parameters, 2) retainable rectification distortion, and 3) robustness against outliers.

2. Preliminary Background

In this section, we briefly describe the relevant theoretical background for image sectification [6,8,9].

Let ($\mathbf{P}_{ol}$, $\mathbf{P}_{or}$, $\mathbf{P}_{nl}$, $\mathbf{P}_{nr}$) denote the original left, origin right, new left, and new right projection matrices, respectively. The relation between the original and new projection matrices in terms of the rectification homography $\mathbf{H}$ can be written as $\mathbf{P}_n$=$\mathbf{H}\mathbf{P}_o$. For the metric projection matrix, represented as $\mathbf{P}$=$\mathbf{K}[\mathbf{R}^{\mathrm{T}} \mid -\mathbf{R}^{\mathrm{T}}\,\mathbf{t}]$, where $\mathbf{K}$, $\mathbf{R}$, and $\mathbf{t}$ denotes the calibration matrix, rotation matrix, and translation vector, respectively, the rectification homography can be written as follows:

$$\mathbf{H} = \tilde{\mathbf{P}}_n \tilde{\mathbf{P}}_o^{-1} = \mathbf{K}_n \mathbf{R}_n \mathbf{R}_o^{-1} \mathbf{K}_o^{-1} = \mathbf{K}_n \mathbf{R}' \mathbf{K}_o^{-1} \qquad (1)$$

where $\tilde{\mathbf{P}}$ denotes the left 3×3 sub-matrix of $\mathbf{P}$ and $\mathbf{R}' = \mathbf{R}_n \mathbf{R}_o^{-1}$ is the combined rotation matrix. Based on the two-view epipolar geometry, i.e. $\mathbf{m}_r{}^T \mathbf{F} \mathbf{m}_l = 0$ for a pair of corresponding image points ($\mathbf{m}_l$, $\mathbf{m}_r$) and $\mathbf{F}$ is the associated fundamental matrix, this linear constraint on the rectified image coordinates can be formulated as:

$$\left(\mathbf{H}_r \mathbf{m}_r^j\right)^T [\mathbf{u}_1]_\times \left(\mathbf{H}_l \mathbf{m}_l^j\right) = 0 \qquad (2)$$

where the 3-vector $\mathbf{u}_1 = (1,0,0)^T$, and $[\ ]_\times$ denotes a 3×3 skew symmetric matrix defined as a cross product operator of two 3-vectors, i.e. $[\mathbf{a}]_\times \mathbf{b} = \mathbf{a}\times\mathbf{b}$ [1], the index *r* and *l* denote the right and left images, and the index j denotes the *j*-th pair of correspondence points. Note that $[\mathbf{u}_1]_\times$ is a specific form of the fundamental matrix for a rectified image pair. In [6], the authors used equation (2) as the objective function, parameterized by $\mathbf{H}_r$ and $\mathbf{H}_l$ with 10 d.o.f (2 for $\mathbf{H}_r$ and 8 for $\mathbf{H}_l$), for the cost minimization on manually selected match points.

From equation (1) and (2), the fundamental matrix can be re-written as follows:

$$\begin{aligned}\mathbf{F} &= \mathbf{H}_r^T [\mathbf{u}_1]_\times \mathbf{H}_l \\ &= \mathbf{K}_{or}^{-T} \mathbf{R}_{or}^{-T} \mathbf{R}_{nr}^T \mathbf{K}_{nr}^T [\mathbf{u}_1]_\times \mathbf{K}_{nl} \mathbf{R}_{nl} \mathbf{R}_{ol}^{-1} \mathbf{K}_{ol}^{-1} \\ &= \mathbf{K}_{or}^{-T} \mathbf{R}_r'^T \mathbf{K}_{nr}^T [\mathbf{u}_1]_\times \mathbf{K}_{nl} \mathbf{R}_l' \mathbf{K}_{ol}^{-1}\end{aligned} \qquad (3)$$

Note that the calibration matrix $\mathbf{K}$ has 5 d.o.f. and the combined rotation matrix $\mathbf{R}' = \mathbf{R}_o \mathbf{R}_n^T$ has 3 d.o.f. In this formulation, the fundamental matrix is parameterized in terms of the metric camera matrices. In [9], the authors showed that $\mathbf{K}_{nr}^T [\mathbf{u}_1]_\times \mathbf{K}_{nl}$ equals (up to a scale) to $[\mathbf{u}_1]_\times$ when the second and third rows of $\mathbf{K}_{nl}$ and $\mathbf{K}_{nl}$ are chosen the same. In addition, they assumed $\mathbf{K}_{ol} = \mathbf{K}_{or}$ and they were parameterized by a single variable, i.e. focal length. Hence, the fundamental matrix with respect to the Quasi-Euclidean [9] has 7 d.o.f., i.e. 1 for ($\mathbf{K}_{ol}$, $\mathbf{K}_{or}$), 3 for $\mathbf{R}_l'$, and 3 for $\mathbf{R}_r'$.

3. Proposed Method

Inspired by the specific form of fundamental matrices for the rectified coordinates, which remarkably avoids the over-fitting problem in the projective space, we further generalize the image rectification framework to video sequences. Considering the temporal variations and the constraints on the intrinsic parameters, the objective function across all frames can be formulated as follows:

$$E = E_F + \lambda_K E_K + \lambda_t E_t \qquad (4)$$

where E_F , E_K , E_t represent the spatial error energy, internal energy, temporal smoothness energy, respectively, which will be explained subsequently, and (λ_K, λ_t) are the weights used to balance these three energy terms.

The first energy denotes the spatial error cost, which sums the epipolar constraint errors for all the correspondence points in the left and right images via the fundamental matrix $\mathbf{F}$, is given by,

$$E_F\left(K_{ol}^{(t)}, K_{or}^{(t)}, R_{ol}'^{(t)}, R_{or}'^{(t)}\right) = \sum_j \rho_s\left(f_{Samp}\left(\mathbf{F}^{(t)}, \mathbf{m}_l^{j(t)}, \mathbf{m}_r^{j(t)}\right)\right) \qquad (5)$$

The function f_{Samp} is defined as the Sampson error [1] for the j-th pair of correspondence points associated with the fundamental matrix $\mathbf{F}$, given as follows:

$$f_{Samp}\left(\mathbf{F}, \mathbf{m}_l^j, \mathbf{m}_r^j\right) = \frac{\left\|\mathbf{m}_r^{jT} \mathbf{F} \mathbf{m}_l^j\right\|^2}{\left\|\tilde{\mathbf{I}}_3 \mathbf{F} \mathbf{m}_l^j\right\|^2 + \left\|\tilde{\mathbf{I}}_3 \mathbf{F}^T \mathbf{m}_r^j\right\|^2} \qquad (6)$$

where the matrix $\tilde{\mathbf{I}}_3 = diag(1,1,0)$ is used to indicate the first two entries of the 3-vector. In eq. (5), the robust error function ρ_s is employed to alleviate the influence of outliers, which is given by $\rho_s(r) = \log(1 + r^2 / 2\hat{\sigma}^2)$, known as Lorentzion (or Cauchy) function in robust statistics [10]. Note that the robust standard deviation $\hat{\sigma}$ is

Figure 2. The rectification results on the *hiking* sequence by using the proposed method. Two columns are the left and right views.

self-determined by the order statistics method [11].

The second energy E_K denotes the constraints on the intrinsic parameters in the corresponding camera matrices. The intrinsic parameters of the camera matrix have some reasonable ranges and relations [12], and the imposed constraints on them, e.g. identical focal lengths, zero skew, and principle points close to image center. These constraints are formulated as soft constraints in the function f_K, thus the energy E_K is defined as

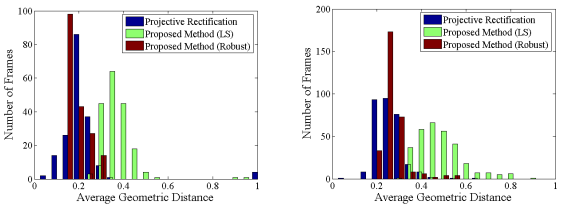
$$E_K\left(K_{ol}^{(t)}, K_{or}^{(t)}\right) = f_K\left(K_{ol}^{(t)}\right) + f_K\left(K_{or}^{(t)}\right) \tag{7}$$

The last energy E_t denotes the temporal smoothness constraints on varying intrinsic and extrinsic parameters and is formulated as:

$$\begin{aligned} &E_t\left(K_{ol}^{(t-1)}, K_{or}^{(t-1)}, R_{ol}^{\prime(t-1)}, R_{or}^{\prime(t-1)}, K_{ol}^{(t)}, K_{or}^{(t)}, R_{ol}^{\prime(t)}, R_{or}^{\prime(t)}\right) \\ &= f_{\delta K}\left(K_{ol}^{(t-1)}, K_{ol}^{(t)}\right) + f_{\delta K}\left(K_{or}^{(t-1)}, K_{or}^{(t)}\right) \\ &+ f_{\delta R}\left(R_{ol}^{\prime(t-1)}, R_{ol}^{\prime(t)}\right) + f_{\delta R}\left(R_{or}^{\prime(t-1)}, R_{or}^{\prime(t)}\right) \end{aligned} \tag{8}$$

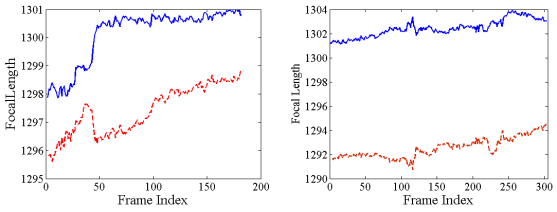
where the functions $f_{\delta K}$ and $f_{\delta R}$ are used to penalize the variations between the current and previous time states.

The Levenberg-Marquardt optimization algorithm with box constraints [13] is applied to solve the nonlinear optimization problem in equation (4). For the first frame, we zero the rotation along the z-axis on the left image to suppress the rectification distortion as well as reduce the ambiguities in the solution. For the subsequent frames, the parameters are initialized by the results of the previous time state.



(a) (b)

Figure 3. The histogram of geometric errors in the (a) *walk away* sequence and (b) *hiking* sequence.



(a) (b)

Figure 4. The curves of the estimated focal lengths in the (a) *walk away* sequence and (b) *hiking* sequence. The blue curves and red dash curves indicate the left and right cameras, respectively.

4. Experimental Results

In our implementation, the box constraints for the first frame are set to within ± 15 degrees in Euler angles and [-1,1] for the focal length parameter α, where $f = 3^{\alpha} \cdot (w+h)$ for image width w and height h [12]. The initial values of all the parameters are set to be zero. For the subsequent frames, the initial values are set to the previous results, and the box constraints for the rotation angles are set to within ± 5 degrees of deviation in Euler angles. We use the above setting for all the experiments in this paper.

All the experiments were performed on the PC with Intel Core2 CPU 6320 of 1.86 GHz 1.87GHz, and DDR RAM 2G. The corresponding feature points were automatically detected and matched via the SIFT feature extraction and matching [14].

We show the experimental results on two stereo sequences, i.e. the *walk away* sequence, as shown in Figure 1, and the *hiking* sequence, depicted in Figure 2. Figure 1 depicts the problem of the conventional projective rectification due to the over-fitting problem in projective space and the significant rectification distortion.

The efficiency of the proposed method is summarized as follows. For the first frame, we set the maximal number of the Levenberg-Marquardt iterations to 200, and it terminated in 2 seconds. For the subsequent frames, the maximal iteration number was set to 20, and it executed at the rate of 2.5-3 fps, depending on the number of match points. Note that the execution time does not include other operations, such as the feature extraction/ matching and image warping.

The histogram of the geometric error distribution is shown in Figure 3. The error measures the average distances of the parallel line of the feature point and its corresponding point on the rectified image pairs. In this figure, we observe the geometric errors of the projective rectification are generally smaller than those of the pro-