A data-fusion approach to motion-stereo

Francesco Malapelle[a,*], Andrea Fusiello[a], Beatrice Rossi[b], Pasqualina Fragneto[b]

[a] *Dipartimento di Ingegneria Elettrica, Gestionale e Meccanica, University of Udine, Via Delle Scienze, 208 - 33100 Udine, IT*

[b] *AST Lab - STMicroelectronics Via Camillo Olivetti, 2 - Agrate Brianza (MB), IT*

Abstract

This paper introduces a novel method for performing motion–stereo, based on dynamic integration of depth (or its proxy) measures obtained by pairwise stereo matching of video frames. The focus is on the data fusion issue raised by the motion–stereo approach, which is solved within a Kalman filtering framework. Integration occurs along the temporal and spatial dimension, so that the final measure for a pixel results from the combination of measures of the same pixel in time and whose of its neighbors. The method has been validated on both synthetic and natural images, using the simplest stereo matching strategy and a range of different confidence measures, and has been compared to baseline and optimal strategies.

Keywords: motion-stereo, temporal-stereo, dynamic-stereo, data fusion, Kalman filter, parallax.

1. Introduction

This paper deals with the problem of *motion-stereo*, i.e., depth estimation in a monocular sequence of images taken by a moving camera [31]. Whereas in binocular stereo two cameras separated by a fixed baseline are employed, in motion-stereo a single camera moves through a static scene. As a result, over a period of time, the camera traverses a "baseline" of undetermined length. The grounds for addressing such problem lie in the attempt to solve the *accuracy-precision* trade-off in stereo matching, which can be summarized as follows: due to quantization errors, the estimated disparity is more precise with a larger baseline, but the matching is less accurate, because the of the exacerbation of perspective and radiometric nuisances that cause false and missing matches. There is manifestly a conflict between accuracy and precision, which motion-stereo approaches attempt to reconcile.

Early work in motion-stereo [28, 16, 22], integrates depth maps from different frames into a single map. They require motion and camera parameters to be known, and most of them restricts to lateral motion. A common drawback is that they warp the disparity map from frame to frame, thereby introducing errors and approximations that disrupt the prediction, and make the integration pointless. More recent motion-stereo approaches aggregate measures in a discretized 3D volume [30, 18, 32], but they need calibrated cameras as well.

The multiple-baseline approach [13, 19, 11] generalizes binocular stereo by computing an aggregated matching cost which considers all the images simultaneously, and then proceeds as in the binocular case. These methods require camera centers to be collinear (equivalent to lateral motion). Generalizations of these approaches can be found in the multi-view stereo literature, where the aggregated cost is computed along the optical ray in a discretized volume [7, 6].

From the geometrical point of view, the problem raised by motion-stereo is how to set a common reference frame where measures from different images can be integrated. The discretized volume seems the natural choice, however computation in 3D space can be avoided by considering image-based quantities such as depth, binocular disparity or *planar parallax*. It will be shown in Section 2 that when camera parameters and its motion are unknown, planar parallax is a suitable depth–proxy that generalizes disparity and depth. This approach based on pixel-based measures – also called "iconic" – is motivated by applications like view synthesis, video interpolation and enhancement (frame rate up-conversion) and free viewpoint 3D TV.

In this work *we concentrate on the data fusion problem posed by the motion–stereo approach*, being agnostic with respect to: i) the depth–proxy that is being used ii) the binocular stereo matching algorithm, which is considered as part of the input of our method. As in [16, 28], we use a *dynamic* approach, as we apply Kalman filtering for recursive estimation of depth maps by combining measurements along the time line and within a spatial neighborhood.

*Corresponding author

Email addresses: name.surname@uniud.it (Francesco Malapelle), name.surname@uniud.it (Andrea Fusiello), name.surname@st.com (Beatrice Rossi), name.surname@st.com (Pasqualina Fragneto)

Pixel–wise depth measures are relaxed by considering the information coming from the neighbors within the same *superpixels*, using a spatial Kalman filter. In both temporal and spatial dimensions, the depth measures are trusted using confidence metrics attached to the measures.

An analogous result has been obtained in [16] by smoothing disparity maps with piecewise continuous splines, where a regularization-based smoothing is used to reduce measurement noise and to fill in areas of unknown disparity. Other methods perform adaptive smoothing in a *edge–aware fashion*, e.g. [12] where temporal consistency is enforced among different depth maps using an edge-aware Gaussian filtering extended to the temporal dimension in video volumes, or [25] where the depth map is filled by solving a least square error problem using edge and temporal information as weights. With respect to our approach, the key difference is that these works are post–processing approaches that aim at improving the quality of depth maps whereas our method uses edge information (in the form of superpixels) to be aware of which neighbors are relevant *while* updating depth values on the current reference map.

A preliminary version of this work appeared in [14] without the spatial relaxation.

Contribution

The main contribution of this paper is a data-fusion framework for motion-stereo, integrating both temporal and spatial information. Also, a comprehensive review of the available depth-proxies is presented in a unified framework and it is shown how planar parallax can be applied with general motion and unknown camera parameters.

The paper can be seen as a general-motion, uncalibrated extension of a classical work [16], which constrained motion to be lateral and required camera internal parameters. Moreover, [16] warped the disparity map from frame to frame, thereby introducing errors and approximations that disrupted the prediction (as shown by our experiments), whereas we fix this by keeping the reference frame constant.

We also report an extensive comparison of several confidence measures in the context of our approach.

Paper Structure

This paper is structured as follows: in Section 2 we survey some background knowledge, in particular we present three suitable candidates for the depth–proxy. In Section 3 we present our method: stereo processing is described in Section 3.1, which produces the input data for the subsequent step in the form of depth measures. Then, the actual core of the algorithm (described in Sections 3.2 and 3.3) merges input measurements into the final result. In Section 4 we report experimental results and we draw conclusions in Section 5.

2. Background: depth–proxies

We do not make any hypothesis on whether the camera is calibrated or not, or if motion is constrained/known or not. These assumptions affects the choice of the *depth–proxy*. Several depth–proxies can be computed depending on factors such as the constraints on the motion of the camera and/or the availability of the perspective projection matrices. The depth–proxy must depend only on the reference frame and not on the other frames being considered. In this way each iteration provides a new estimate commensurate with the others. In this section we present three suitable candidates.

2.1. Depth

The depth of a point is its distance from the focal plane of the camera. If the interior camera parameters are available, stereo correspondences can be converted directly into depth values. The depth values for a given pixel obtained from subsequent frames are directly comparable.

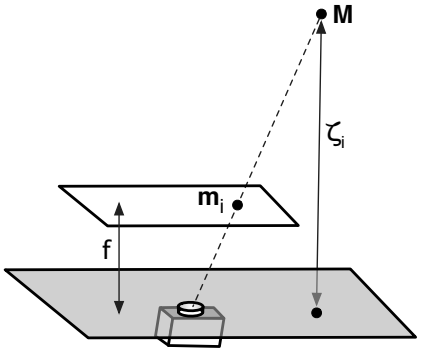

Figure 1: The depth ζ_i is the distance of the 3D point from the focal plane of the camera (shaded in the picture).

Let $\mathbf{M}$ be a 3D point and let $(\mathbf{m}_r, \mathbf{m}_i)$ be its projections onto the image planes I_r and I_i respectively. Let $P_r = K_r[R_r|\mathbf{t}_r]$ and $P_i = K_i[R_i|\mathbf{t}_i]$ be the perspective projection matrices of the two cameras (that must be known). The equation of the epipolar line of $\mathbf{m}_r$ in I_i is

$$\zeta_i \mathbf{m}_i = \mathbf{e}_i + \zeta_r K_i R_i R_r^\top K_r^{-1} \mathbf{m}_r \tag{1}$$

where $\mathbf{e}_i := K_i(t_i - R_i R_r^\top t_r)$ is the epipole and ζ_r and ζ_i are the unknown depths of $\mathbf{M}$ (with reference to P_r and P_i, respectively). Thus we can write

$$\mathbf{e}_i = \zeta_i \mathbf{m}_i - \zeta_r \mathbf{m'}_r \tag{2}$$

where $\mathbf{m}'_r := K_i R_i R_r^\top K_r^{-1} \mathbf{m}_r$. Since the three points $\mathbf{e}_i$, $\mathbf{m'}_r$ and $\mathbf{m}_i$ are collinear, one can solve for ζ_r using the following closed form expression [10]

$$\zeta_r = \frac{(\mathbf{e}_i \times \mathbf{m}_i)(\mathbf{m}_i \times \mathbf{m'}_r)}{\|\mathbf{m}_i \times \mathbf{m'}_r\|^2}. \tag{3}$$

Since in real situations camera parameters and image locations are known only approximately, the back–projected rays do not actually intersect in space. However, it can be shown [10] that Formula (3) solves Equation (2) in a least squares sense.

The actual computation of depth values is performed by applying Equation (3): $\mathbf{e}_i$ is obtained as the projection of the optical center of the reference camera C_r, through the second camera P_i; the set of dense correspondences $(\mathbf{m}_r^k; \mathbf{m}_i^k)$ with $k = 1, \ldots, K$, where K is the number of correspondences for the current image pair, is known from the stereo matching step; image points $\mathbf{m}_i'$ are computed according to Equation (2).

Please observe how this formulation elegantly avoids the explicit triangulation of $\mathbf{M}$, which would be required in a naive approach.

2.2. Disparity

If interior camera parameters are unavailable, the binocular disparity is the first depth–proxy that is readily available from stereo correspondences. However, the disparity values of a pixel computed from subsequent frames are commensurate only if motion is constrained such that all cameras share a common focal plane (the focal plane is parallel to the image plane and contains the camera center).

When two focal planes are coplanar (i.e. up to coordinate change, motion is along X axis) then $\zeta_i = \zeta_r := \zeta$ and the epipole is $\mathbf{e}_i = \begin{bmatrix} b_i f & 0 & 0 \end{bmatrix}^\top$, where f is the focal length b is the magnitude of the translation. Moreover, if $K_i = K_r$ then $\mathbf{m'}_r = \mathbf{m}_r$, hence Equation (2) simplifies to:

$$\mathbf{m}_i - \mathbf{m}_r = \begin{bmatrix} b_i f/\zeta & 0 & 0 \end{bmatrix} \tag{4}$$

This configuration is also called *normal* case (for stereo). The disparity, defined only in the normal case, is the nonzero (horizontal) component of the pixel coordinates differences. Two cameras can be always brought to the normal case by rectification [5, 4].

In the case of multiple cameras, since disparity is proportional to the reciprocal of the depth and the depth is defined with respect to the focal plane, there must be a common focal plane in order for disparities to be commensurate. This can always be achieved for $N \leq 3$ cameras by rectification (rotating the focal planes around the optical centers until they coincide with the plane defined by the three centers), but cannot be guaranteed for more cameras, unless camera centers lies on a plane.

2.3. Planar Parallax

In the case where camera calibration is unavailable and the camera undergoes a general motion, *planar parallax* can be profitably employed instead of depth. *Planar parallax* represents the displacement in the apparent position of objects imaged from different points of view with respect to a reference plane [23], and can be computed from stereo correspondences.

In this section we review some background notions needed to understand the proposed methodology. A complete discussion and formulation of the planar parallax theory can be found in [26, 9].

Let us consider a 3D point $\mathbf{M}$ belonging to some space plane Π and its projection $(\mathbf{m}_r, \mathbf{m}_i)$ onto the image planes I_r and I_i respectively. There exists a non–singular linear transformation, or homography, that maps $\mathbf{m}_r$ onto $\mathbf{m}_i$, that is

$$\mathbf{m}_i \simeq H_\Pi \mathbf{m}_r \tag{5}$$

where H_Π is the homography induced by plane Π and $\simeq$ means equality up to a scale factor. For 3D points $\mathbf{M}$ not belonging to plane Π, the following more general relation holds:

$$\mathbf{m}_i \simeq H_\Pi \mathbf{m}_r + \mathbf{e}_i \gamma \tag{6}$$

where $\mathbf{e}_i$ is the epipole in I_i and γ is the *planar parallax* (or, simply, *parallax* if the context is clear), which can be interpreted as the displacement between the point $H_\Pi \mathbf{m}_r$ mapped via the homography H_Π and its actual corresponding point $\mathbf{m}_i$.

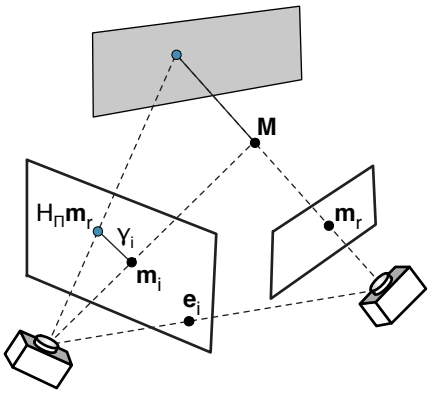

Figure 2: The parallax γ_i is the length of the segment joining $\mathbf{m}_i$ and $H_\Pi \mathbf{m}_r$. The reference plane Π is shaded in the picture.

Given point correspondences and a plane homography H_Π, parallax values can be obtained for each pixel of the reference frame by solving for γ in Equation (6):

$$\frac{1}{\gamma} = \frac{(\mathbf{e}_i \times \mathbf{m}_i)^T (\mathbf{m}_i \times H_\Pi \mathbf{m}_r)}{\|\mathbf{m}_i \times H_\Pi \mathbf{m}_r\|^2}. \tag{7}$$

One can contrast this equation with Equation (3) and observe that they coincides if $\mathbf{m}_r' = H_\Pi \mathbf{m}_r$, in which case $\frac{1}{\gamma} = \zeta_r$. In particular, it can be seen that this condition is equivalent to the special choice $H_\Pi = H_\infty$, where H_∞ is the infinite plane homography, i.e. the homography induced by the infinite plane between the pair of images (I_r,I_i).

Furthermore in the stereo normal case then H_∞ is the identity and the epipole is $\mathbf{e}_i \simeq \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}^\top$, thus parallax in Equation 6 results to be proportional to binocular disparity.

To summarize: when $H_\Pi = H_\infty$ the parallax γ reduces to the reciprocal of the depth (while in general it is proportional to it), and in the normal case it is proportional to disparity. Moreover, it can be demonstrated that γ depends only on the reference image and the plane Π, and not on the parameters of the second image. This is why the parallax can be seen as a useful generalization of the depth and (inverse) disparity.

By setting the reference image, together with a fixed reference plane Π, one can thus obtain a projective proxy for the depth of a point that is consistent across several images of a same, modulo a global scale factor. In fact, independent estimates of parallax derived from different image pairs (I_r, I_i), $i = 2, \ldots, N$ differ from each other by an unknown scale factor, which must be estimated independently.

In practice, parallax values are computed using (7) for each pixel: as for the depth, the dense set of correspondences $(\mathbf{m}_r^k; \mathbf{m}_i^k)$ on the pair of images (I_r, I_i) is known from the stereo matching step; the homography H_{ri}^Π is obtained according to the procedure explained in [14] and epipole $\mathbf{e}_i$ is estimated from epipolar geometry.

Finally, we saw that Equation (6) describes the relationship between two views through a reference plane. Since γ does not depend on the position of the second camera, we can replace the second image with a *new one*, thus we can *transfer* or *warp*, pixel $\mathbf{m}_r$ onto $\mathbf{m}_n$ with:

$$\mathbf{m}_n \simeq H_\Pi \mathbf{m}_r + \mathbf{e}_n \gamma \tag{8}$$

where H_Π and $\mathbf{e}_n$ define the position of the new camera. This can be used to transfer a parallax map from one reference frame to another. This operation brings in several issues related to non-injectivity and non-surjectivity of the transfer map, that are well known in the context of view-synthesis [17].

3. Proposed Method

The input of the method is a monocular video sequence of N frames, of which one is set as the reference, denoted by I_r. For every pair of images (I_r, I_i) (where, for example, if $I_r = I_1$ and $i = 2, \ldots, N$), estimates of the depth–proxy map relative to the reference frame are computed independently by binocular stereo matching.

We designate parallax as the depth–proxy, for it is the more general one and subsumes all the others. However disparity or depth can be used instead when certain conditions are fulfilled. Regarding camera motion, the only assumption made is that a relevant portion of the reference frame is kept visible at all the subsequent frames of the video segment. When this assumption fails, a new reference frame is set and the filter is restarted. Note that information about the temporal trajectory is not used, i.e. the pairs could be processed in any order. This property has two main advantages: i) pairwise processing can be performed independently, making the algorithm highly parallelizable ii) sets of still images as input, instead of video sequences, can be processed.

Each of the $N-1$ independent estimates of the parallax map contains errors and valuable information: the goal of the data fusion is to enhance the latter while smoothing out the former. In our framework all these parallax maps are combined together using spatial and temporal coupled Kalman filters, achieving more stable and accurate values. Superpixels provide the spatial support for the relaxation of parallax values among the image neighbors.

The rationale behind motion-stereo is to break the accuracy vs precision trade-off by using multiple baseline lengths: a small baseline implies few occlusions, easier stereo matching but raw quantization of the parallax, whereas a large baseline implies better quantization of the parallax but more occlusions and harder matching.

3.1. Stereo Matching

The image pairs $(I_r,\ I_i)$ needs to be rectified for the subsequent stereo matching step to work. In particular, each pair must be rectified independently, unless the camera centers are coplanar. In the calibrated scenario, we use [5], where the algorithm takes the perspective projection matrices of the original cameras and computes a pair of rectifying projection matrices. When internal parameters are unknown, we use [4], which assumes that a number of corresponding points are available and we seeks the rectifying homographies that make the original points satisfy the epipolar geometry of a rectified image pair.

Dense correspondences between I_r and I_i can be obtained using any stereo matching algorithm. In our experiments, since we focus on the *integration framework* and not on the performance of the stereo itself, we used a simple block–matching with Normalized Cross Correlation (NCC) as a matching score:

$$\frac{\sum_{n \in W} (I_r(x_n, y_n) - \mu_r)(I_i(x_n, y_n) - \mu_i)}{\sqrt{\sum_{n \in W} (I_r(x_n, y_n) - \mu_r)^2} \sqrt{\sum_{n \in W} (I_i(x_n, y_n) - \mu_i)^2}} \tag{9}$$

where μ_r and μ_i are the averages of window Win images I_r and I_i, respectively.

After the block–matching step, we perform a left–right consistency (LRC) check, which is a standard procedure based on the uniqueness principle [15]. The consistency is verified if p is matched with p' when searching on the pair $(I_r,\ I_i)$ and p' is matched with p when searching on the pair $(I_i,\ I_r)$, where p is a point in I_r and p' is a point in I_i. All non–consistent matches are discarded. This procedure skims the results from occluded pixels and bad matches. Dense correspondences are then transferred back to the original reference images by applying the inverse of the rectifying homographies (de–rectification).

During the stereo matching step, a confidence map, associated to the parallax map, is also computed. For

each pixel we integrate the LRC check with a confidence indicator based on the matching score profile.

Thus, the confidence associated to the parallax computed at pixel i is,

$$\varphi(i) := \begin{cases} 0 & \text{if pixel } i \text{ fails the LRC check} \\ \phi_*(i) & \text{o/w} \end{cases} \tag{10}$$

where $\phi_*(i)$ is one of the confidence metrics discussed in Appendix A. The confidence $\varphi(i)$ varies in $[0, 1]$, where 0 means that pixel i is *totally unreliable* and 1 means *maximally confident*.

3.2. Temporal integration

Temporal integration of parallax data is performed through a simple 1-d Kalman filter with constant (up to a scale) state and direct measurement model. Let $x_t(m)^+$ be the best parallax estimate (the state) available at time t for pixel m, let $p_t(m)^+$ be its variance; let $z_t(m)$ be the parallax measured at pixel m of frame t (via stereo matching), and let $r_t(m)$ be its variance. The Kalman filter equations write:

$$\textbf{Process:}\quad x_t = s \cdot x_{t-1} + w_t \quad \mathrm{Var}(w_t) = q_t \tag{11}$$

$$\textbf{Measure:}\quad z_t = x_t + v_t \quad \mathrm{Var}(v_t) = r_t \tag{12}$$

$$\textbf{Prediction:}\quad x_t^- = s \cdot x_{t-1}^+ \quad p_t^- = s^2 \cdot p_{t-1}^+ + q_t \tag{13}$$

$$\textbf{Update:}\quad x_t^+ = \frac{x_t^- r_t + z_t\, p_t^-}{p_t^- + r_t} \quad p_t^+ = \frac{p_t^- r_t}{p_t^- + r_t} \tag{14}$$

Where x_t^- and p_t^- represent the *a priori* estimations of the state and its variance respectively, whereas x_t^+ and p_t^+ are their updates using measurement z_t and its variance r_t. The variable m has been omitted as the treatment is uniform over the pixels.

It turns out to be more convenient to formulate the update equations in terms of the inverse variance, which will be henceforth called *information* (the *Fisher information* of a random multivariate distribution is the inverse covariance [3]). Let ${}^ip = 1/p$ and ${}^ir = 1/r$, then Equation (14) becomes:

$$x_t^+ = \frac{z_t\, {}^ir_t + x_t^-\, {}^ip_t^-}{{}^ir_t + {}^ip_t^-} \quad {}^ip_t^+ = {}^ir_t + {}^ip_t^-. \tag{15}$$

The process model contains a multiplicative factor s which takes into account the fact that independent measures of the parallax are scaled by an unknown factor: in fact, the current state is always scaled to match the measure. The scale s is estimated by comparing x_{t-1}^+ with z_t in a robust (outliers resilient) way. First the ratio between the two maps is computed pixelwise, considering only the pixels that, given their information value, are the most reliable (i.e. upper quartile of the ${}^ir_t$ map); then the ratios which are greater than 5.2 median absolute deviations from the median are discarded as outliers (a.k.a. x84 rejection rule [21]); finally the scale is computed as the mean of the inlier ratios.

The process noise w_t accounts for the errors introduced in predicting the state. Since the state we are estimating is constant (up to a scale), and no approximation are made in the prediction, our temporal model has $q_t = 0$.

The measurements noise v_t models errors that affect the parallax estimation, hence its information ${}^ir_t$ is directly related to the confidence φ defined in Equation (10). We use ${}^ir_t = 12\varphi$, which sets the maximum information for a correct parallax value to the reciprocal of the variance of the quantization noise (which is $1/12$).

The update of the filter state takes place through a *validation gate* to ensure that outliers do not skew the estimate. In particular, we consider the Mahalanobis distance as a gating criterion [27]. The update is accepted only if:

$$\frac{(x_t^- - z_t)^2}{p_t^- + r_t} \leq \chi_1^2(\alpha) \tag{16}$$

where $\chi_1^2(\alpha)$ is the upper $100\alpha^{th}$ percentile of a chi-square distribution with 1 d.o.f. (we used $\alpha = 0.98$).

The update equation fails when ${}^ip_t^- = {}^ir_t = 0$, because a 0/0 form is obtained. This happens at $t = 1$ if a reliable measure (${}^ir_1 \neq 0$) is not available, and at any subsequent t until a reliable measure is found. This special case is handled within the validation gate by simply skipping the update whenever ${}^ir_t = 0$. Please note that ${}^ir_t = 0$ means that the pixel is unmatched (not visible in the conjugated image).

In the most general case, the filter starts with ${}^ip_0^- = 0$ and x_0^- undefined, however, if a parallax map is available for the reference frame of the previous video segment, it can be warped to the current reference frame with Equation (8) and provides a partial initialization for the state. The information of the warped parallax is downweighted by a factor 10 to account for errors introduced by the warping.

Finally, it is worth noting here that this simple Kalman filter – ignoring the scale s – reduces to a weighted average of the measures z_t with the information values ${}^ir_t$ as weights, as can be observed by solving the recursive update equations, thus obtaining:

$${}^ip_t^+ = \sum_{k=0}^{t} {}^ir_k \tag{17}$$

$$x_t^+ = \frac{z_t\, {}^ir_t + x_{t-1}^+ \sum_{k=0}^{t-1} {}^ir_k}{\sum_{k=0}^{t} {}^ir_k} = \frac{\sum_{k=0}^{t} z_k\, {}^ir_k}{\sum_{k=0}^{t} {}^ir_k}. \tag{18}$$

Indeed, the middle term of Equation (18) is the well known formula for the recursive computation of the average. A matrix equivalent of Equations (17) and (18) can be also derived as the least squares solution to the problem of optimally (in terms of Mahalanobis distance) combining an ensemble of independent (multivariate) random variables which estimate the same true parameter [20]. The advantage of the Kalman filter is in its recursive formulation, which leads to a *causal* filter that produces at each time

instant (dynamically) the best estimate based on the past measures, whereas the weighted average considers all the measures in a batch.

3.3. Spatial integration

The spatial relaxation requires to identify a neighborhood of each pixel in the reference image where the depth is ideally constant. This is achieved by computing *superpixels*, i.e., compact and almost uniform regions of the image, using the Simple Linear Iterative Clustering algorithm (SLIC) [2], which starts with a regular grid of centers and then locally clusters pixels in the combined five–dimensional color (CIELab) and image coordinates space. The density of the initial grid plus a regularization coefficient are the only two parameters that need to be set. The (approximated) desired size of the superpixels is specified so that

$$\text{number of initial cells} = \frac{\text{reference frame resolution}}{\text{desired size of the superpixel}}.$$

Some segmentation examples are shown in Figure 3.

Once the superpixels are extracted, in principle, the integration with the spatial neighborhood should take place by introducing spatial correlations between neighboring pixels, which entails a state vector of the size of the image (M) and a non-diagonal ($M \times M$) covariance matrix. However, this would become too computationally demanding, so we approximate its effect by modifying the prediction step of the temporal 1-d Kalman filter, without changing its structure. In particular, in the prediction formula (13), we substitute the state x^+_{t-1} with a smoothed state $\hat{x}^+_{t-1}$ that depends on the neighboring pixels *within the same superpixel* (and the information ${}^i p^+_{t-1}$ accordingly).

To be consistent with the temporal dimension, we derive $\hat{x}^+_t$ within the Kalman filter framework. As mentioned in [16], an alternative approach to the prediction of state variance is the so called "exponential age-weighting" of measurements, where the current variance is inflated by a small multiplicative factor [3]:

$$p^-_t = (1+\epsilon) p^+_{t-1}. \tag{19}$$

Equations (17) and (18) can be generalized to:

$${}^i p^+_t = \sum_{k=0}^{t} {}^i r_k \delta^{t-k} \tag{20}$$

$$x^+_t = \frac{\sum_{k=0}^{t} z_k \, {}^i r_k \delta^{t-k}}{{}^i p^+_t} \tag{21}$$

where we introduced $\delta = 1/(1+\epsilon)$ which is the inverse of the exponential age-weighting, since we are dealing with information instead of variance.

These formulae can be translated into the spatial domain by substituting the exponential age-weighting term, which gives smaller weights to older measures, with an *exponential distance-weighing* term (with a parameter $\rho < 1$) which serves the purpose of weighting the measure according to the distance to the current pixel. Let $x(m)^+$ be the parallax (state) at pixel m and let ${}^i p(m)^+$ be its information value:

$${}^i \hat{p}^+(m) = \sum_{q \in \Omega(m)} {}^i p^+(q) \rho^{\|m-q\|} \tag{22}$$

$$\hat{x}^+(m) = \frac{\sum_{q \in \Omega(m)} x^+(q) \; {}^i p^+(q) \rho^{\|m-q\|}}{{}^i \hat{p}^+(m)} \tag{23}$$

where $\Omega(m)$ is the superpixel to which pixel m belongs. In this paragraph we will omit the constant temporal index, as we are dealing with the spatial dimension only.

The information of the combined measure is the *sum* of the information values of the original measures (with exponential distance-weighing), so it is much greater than the original point-wise information. This would be correct only if the combined measures are *not correlated*, but this is not the case here, for neighboring parallax measures are indeed correlated.

The problem of combining correlated measures of the same variable has been addressed in the data fusion literature, and one solution that provides consistent estimates is the Covariance Intersection approach [29], where "consistent" means that the estimated covariance is an upper bound of the true covariance. When considering scalar variables, Covariance Intersection boils down to selecting the measure with the highest information value:

$${}^i \hat{p}^+(m) = \max_{q \in \Omega(m)} \{ {}^i p^+(q) \rho^{\|m-q\|} \} \tag{24}$$

$$\bar{q} = \arg \max_{q \in \Omega(m)} \{ {}^i p^+(q) \rho^{\|m-q\|} \}$$

$$\hat{x}^+(m) = x^+(\bar{q}) \tag{25}$$

Please note that Equation (25) would yield the same value of $\hat{x}^+(m)$ for each pixel $m \in \Omega(m)$ if $\rho = 1$, whereas with $\rho < 1$ it produces different values within the same superpixel. The value of ρ can be computed as a function of the cut-off radius θ (in pixels) at which the function $\rho^{\|m-q\|}$ falls below a given threshold, 10^{-2} in our implementation. The value of θ should be of the order of the stereo matching window size. Please note that, as the smoothing is limited within the superpixel, there is no point in choosing θ larger than the superpixel radius.

The following Matlab pseudo-code illustrates one iteration of the filter: the function takes in input the current state estimate `(x,ip)` and the measure `(z,ir)` and updates the state estimate accordingly. This also shows how temporal and spatial integration are iterated.

```
function [x,ip] = STKalmanStep(x,ip,z,ir)
% update state (x,ip)
% in the face of measure (z,ir)

% prediction
s=compute_scale(x,z);
x=s*x;
ip=1/s^2 * ip;
```

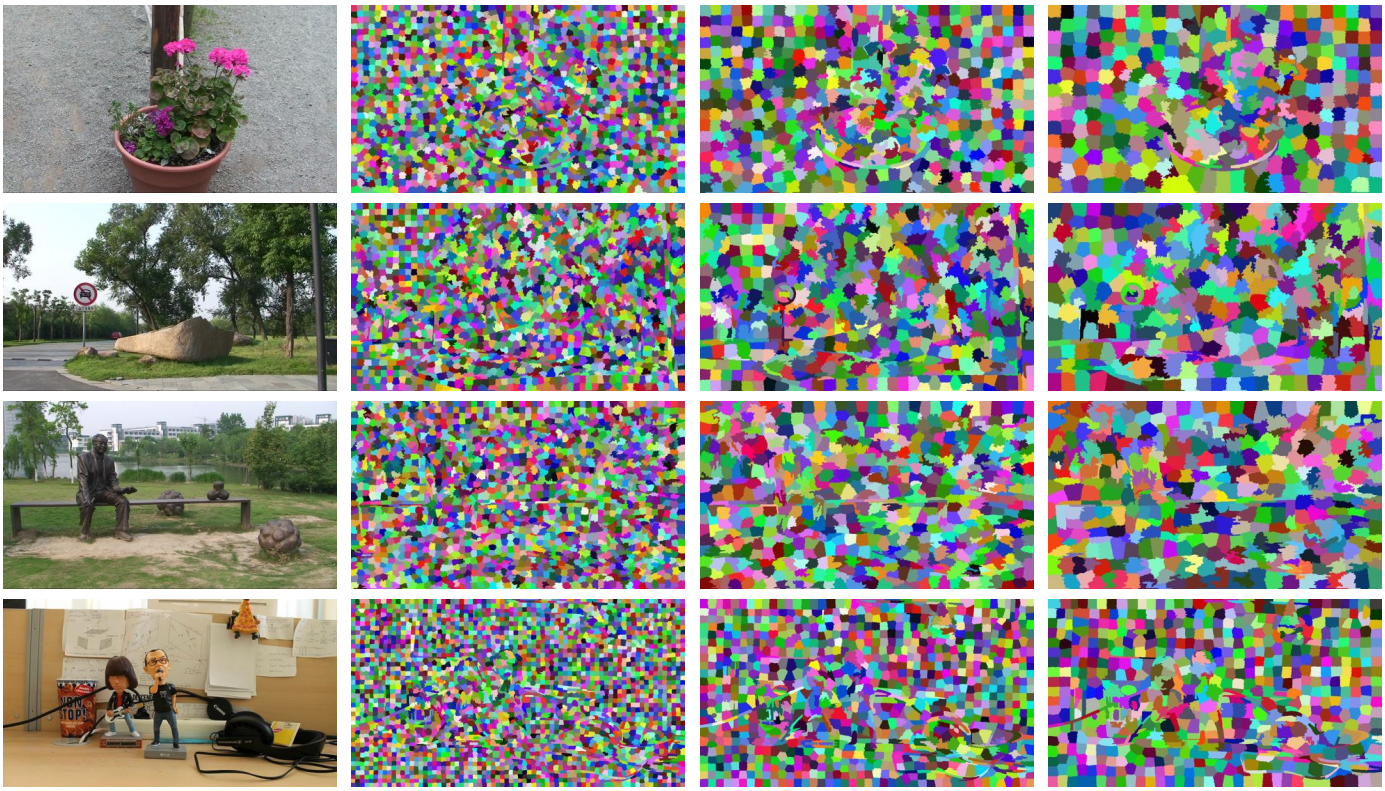

Figure 3: Examples of superpixel extraction using different values for the grid density, which controls the size of the superpixel.

```
% validation gate
res=((x-z).^2)./(1./ip + 1./ir);
v=(res <= Chi) & ir>0;
% temporal update
x(v)=(z(v).*ir(v)+x(v).*ip(v))./(ir(v)+ip(v));
ip(v)=ip(v) + ir(v);
% spatial relaxation
for k=1:numel(superpixels)
    pix=superpixels(k).PixelList;
        for j=1:length(pix)
            w=rho.^sum(sqrt((pix-pix(j)).^2),2);
            [val,pos]=max(ip(pix).*w);
            ip(pix(j))=val;
            x(pix(j))=x(pix(pos));
        end
end
```

The `compute_scale` function implements the robust method described in the text (after Equation (15)). In the `for` cycle we have been sloppy about the difference between linear indexing and subscripts (row, column), for the sake of readability. Also the subtraction in `pix-pix(j)` is not syntactically correct, as `pix(j)` should have been replicated. The actual working code is available on-line [1].

4. Experiments and Results

We run two set of experiments. In the first one we consider images from the Middlebury 2006 datasets [24] with a ground truth in order to validate our method and quantify the benefit of the spatial integration. In the second set we use more general sequences, both outdoors and indoors, without ground truth.

It is important to stress that the method presented here focuses on the fusion of depth measurements, so the results reported should not be evaluated in absolute terms, but relatively to the *input* data, in this case disparity maps produced by NCC block–matching. More sophisticated stereo algorithms coupled with a global optimization yield better depth maps, as those reported, e.g., in [32]. For these reasons a comparison with other stereo methods is pointless, since any of them could be plugged in our framework.

4.1. Middlebury datasets

In the Middlebury datasets the camera motion is constrained along the X axis, so the integration takes place at the disparity level.

The *error rate* is defined as the percentage of computed disparities values whose difference with the ground truth is > 1, as in [24]. Pixels marked as occluded in the ground truth have not been counted.

In all the experiments in this section, we used a square 3×3 window for the NCC stereo matching, and the size of the superpixels is set at 800 pixels.

First we performed a systematic evaluation of the confidence measures described in Appendix A with the Middlebury 2005 datasets. Results are reported in Table 1,

Table 1: Error rates [%] of disparity maps obtained with different confidence measures (see Appendix A for explanation). Best and worst results are highlighted in boldface/green and lightface/red respectively.

Data	GT	MSM	CUR	PKR	MMN	MLM	AML	WMN	UNI
Art	9.58	19.68	20.91	**19.60**	20.11	19.70	19.73	19.67	19.64
Books	8.78	23.14	24.17	22.90	23.51	22.97	23.01	**22.83**	23.03
Dolls	7.95	**16.05**	16.93	16.12	16.69	16.09	16.16	16.16	16.14
Laundry	13.37	28.71	31.73	28.84	30.23	28.96	29.08	**28.37**	29.02
Moebius	8.60	20.34	21.43	20.35	20.64	20.54	20.49	**20.20**	20.36
Reindeer	7.33	14.24	15.60	14.28	15.01	14.23	**14.15**	14.30	14.27
Mean	9.27	20.36	21.79	20.35	21.03	20.41	20.44	**20.25**	20.41

Table 2: Error rates [%] of disparity maps obtained with different confidence measures (see Appendix A for explanation), **without the LRC check**. Best and worst results are highlighted in boldface/green and lightface/red respectively.

Data	GT	MSM	CUR	PKR	MMN	MLM	AML	WMN	UNI
Art	10.61	23.00	25.69	23.19	**22.74**	23.08	23.10	23.27	23.14
Books	11.58	42.90	28.25	25.29	**24.89**	27.42	26.34	25.28	25.53
Dolls	6.29	18.48	20.17	18.62	**18.19**	18.51	18.59	18.61	18.61
Laundry	57.15	44.91	46.94	40.46	**32.39**	44.28	41.33	39.04	41.35
Moebius	6.06	28.42	24.32	22.26	**21.90**	22.70	22.28	22.11	22.20
Reindeer	6.70	18.58	18.95	**16.86**	17.24	22.00	17.91	16.87	16.85
Mean	16.40	29.38	27.39	24.45	**22.89**	26.33	24.92	24.20	24.61

where each entry contains the error rate of the disparity map produced by our method with a given confidence measure.

Table 2 reports the results of a similar experiment in which the confidence measures do not include the LRC, i.e., $\varphi = \phi_*$. These figures compel us to make some observations:

- all confidence measures are equally suited to represent pixel's reliability, for in Table 1 all the entries are very close; however WNM obtains the lowest error rate, probably thanks to the fact that it considers distinctiveness of the match by looking at the second best match, the same recipe that proved so effective in SIFT matching (in fact, PKR, that uses a similar strategy, performs closely to WNM).

- the UNI metric has surprisingly good performances, confirming the robustness of the integration framework; in other words, the data fusion works so well that the confidence becomes nearly irrelevant;

- if LRC is switched off, MMN is the best performer, although by a narrow margin; this suggests that MMN could be a proxy for occlusions detection if LRC cannot be performed;

- the comparison of the two tables indicates that the most important contribution to confidence is the the binary response of the LRC check.

Since WMN obtains the lowest error rate, we chose it as the default confidence measure for the rest of our experiments, although other choices would likely produce similar results.

Then, we assess the benefits of the spatio–temporal integration. Following [8], we consider two touchstones against which to compare the error rate obtained with our method (**Kalman ST**): the **Optimal** map obtained by an oracle that selects the disparity value closest to the ground-truth among all the input estimates for each pixel, and the **Best Map**, obtained selecting the map with the minimum error rate among all the input disparity maps.

Observe that the former represents the theoretical optimum that one can achieve with the given input disparity maps using the temporal dimension, while the latter is an indicator of whether the data-fusion is beneficial with respect to a simple two-views stereo. We also considered other integration strategies: the maximum confidence selection (**Max conf**), which consists in selecting, for each pixel, the disparity that achieves the maximum confidence φ, the temporal fusion (**Kalman T**), that consists in applying only the temporal Kalman filter, without spatial relaxation, as in [14], and our implementation of [16] (henceforth **MKS**), for comparison with another method from

Table 3: Error rates [%] of disparity maps obtained with different fusion strategies (see text for explanation) and WMN as the confidence measure.

Data	Best map	Max conf	Average	MKS [16]	Kalman T	Kalman ST	Optimal
Art	49.76	53.12	48.15	61.79	35.13	**19.67**	21.54
Books	55.89	68.37	59.04	76.89	48.57	**22.83**	29.32
Dolls	42.01	54.71	38.52	57.70	29.01	**16.16**	15.95
Laundry	75.67	69.22	69.27	81.54	58.16	**28.37**	44.23
Moebius	45.73	63.53	45.98	68.19	35.70	**20.20**	22.21
Reindeer	45.11	57.24	49.95	65.05	32.49	**14.30**	17.13
Mean	52.36	61.03	51.82	68.53	39.84	**20.25**	25.06

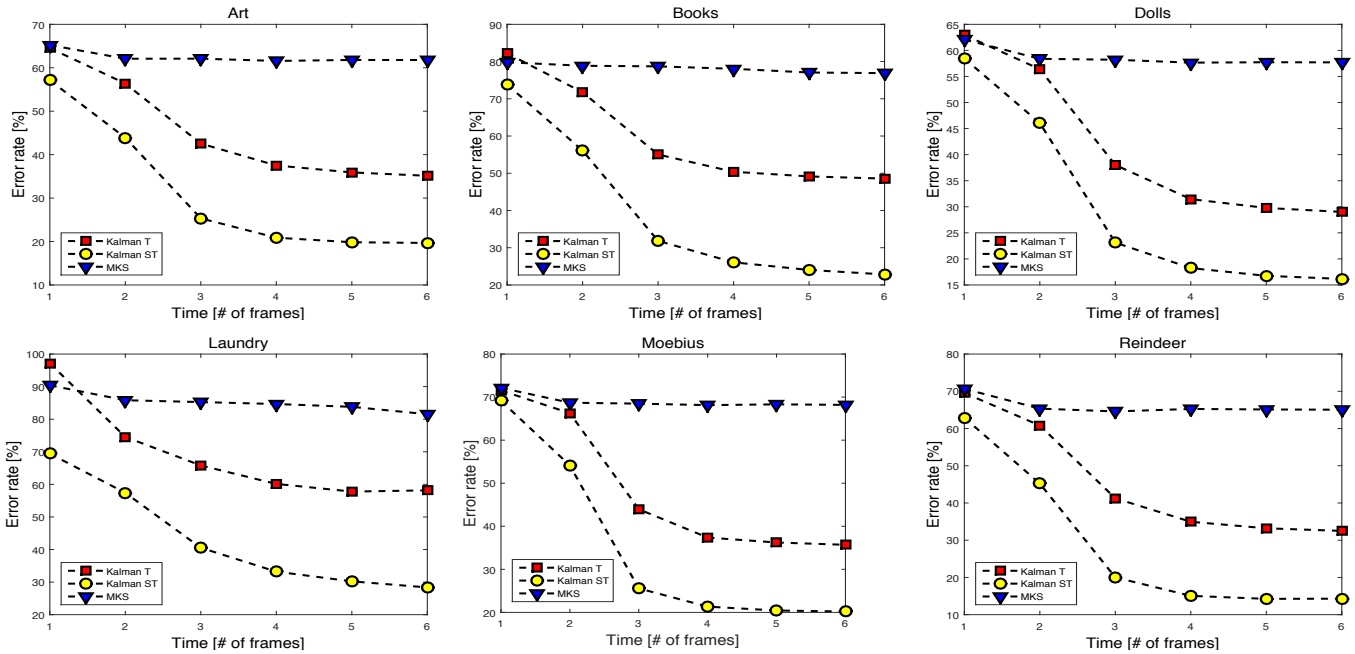

Figure 4: Each graph shows the error rate decreasing as more measures are integrated in the estimation for the three approaches, the **Kalman T** (temporal–only), the **Kalman ST** (spatial support) and the **MKS** (our implementation of [16]).

the literature.

Results with the Middlebury 2005 datasets are reported in Table 3, where it can be appreciated that **Kalman ST** (in boldface) achieves the lowest scores, when compared to the other strategies; in particular spatio–temporal integration always improves the pure temporal Kalman filter and always outperforms the results obtained by **MKS**. Moreover our method always exceeds the best map and, in some cases, it also exceeds the optimal one, due to the spatial relaxation.

Figure 4 reports, for the same experiments, how the error rate decreases as more measures are integrated. Observe that **MKS**, despite the integration and spatial relaxation steps, only slightly improves the results obtained by the regular stereo matching algorithm. This confirms the idea that the warping of the disparity map from frame to frame severely limits the benefits of the integration mechanism. Figures 5 and 6 show qualitative results for the above sequences.

We also considered the 21 sequences of the Middlebury 2006 datasets; results are available on the web [1] and they lead to the same conclusions.

Finally, please note that what we refer to as **Kalman T** is the same implementation of the **Kalman ST** with the spatial step switched-off, which is slightly different from the original one described in [14] because of the validation gate and other tweakings and also because we are using a NCC based stereo algorithm instead of the Census transform. Consequently, figures reported in Tab. 3 are different from those reported in [14].

4.2. Casual video sequences

In the second set of experiments we test the method on the "Flower", "Road", "Lawn", from [32]. These are casual, uncalibrated sequences, hence we used parallax as

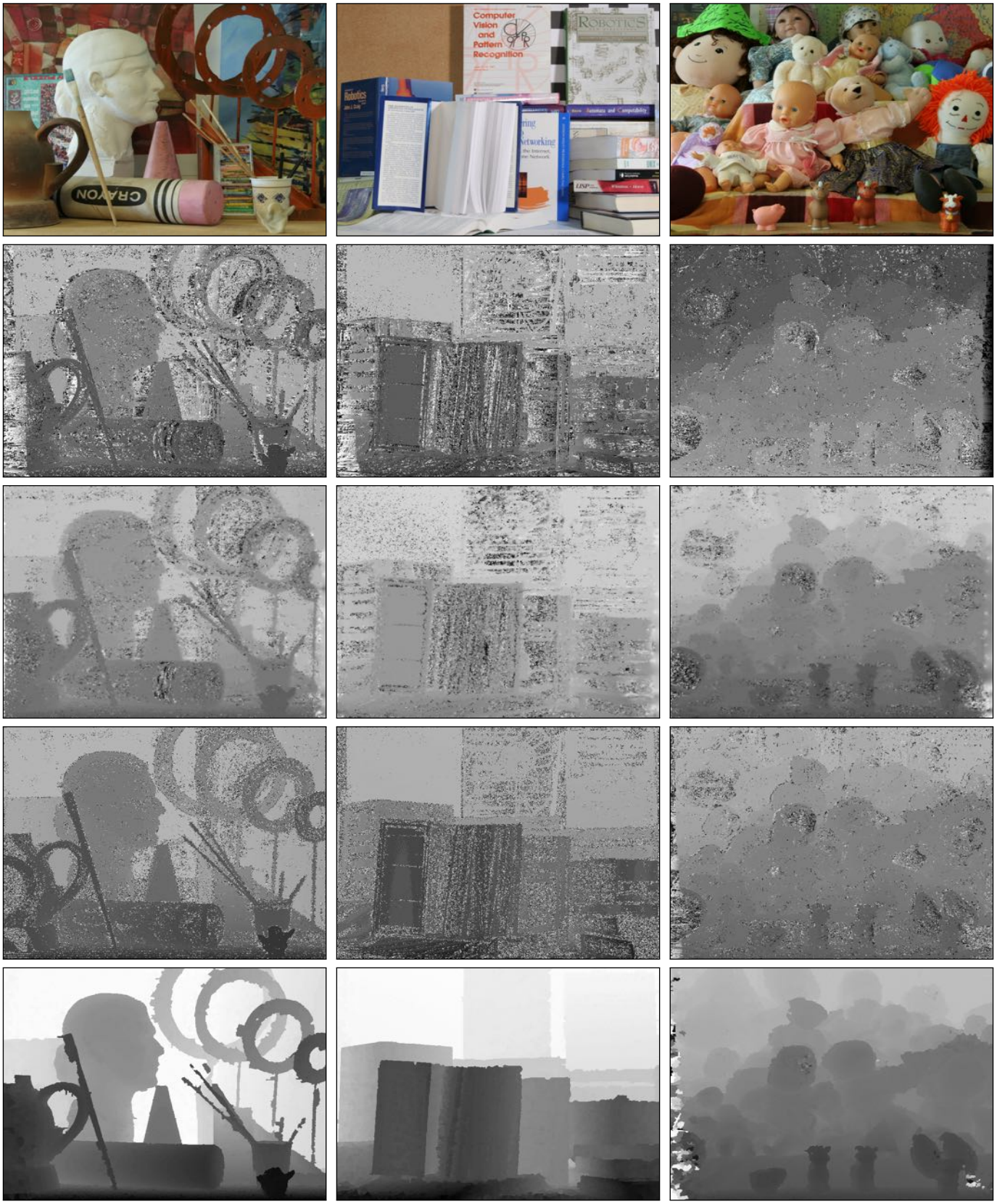

Figure 5: From top to bottom: the reference frame, the **Best Map**, the result of **MKS**, the result of **Kalman T** (temporal–only) and the result of **Kalman ST**. Images are automatically scaled in the range [0,255], hence the gray levels changes from row to row. Full resolution images can be seen on line [1].

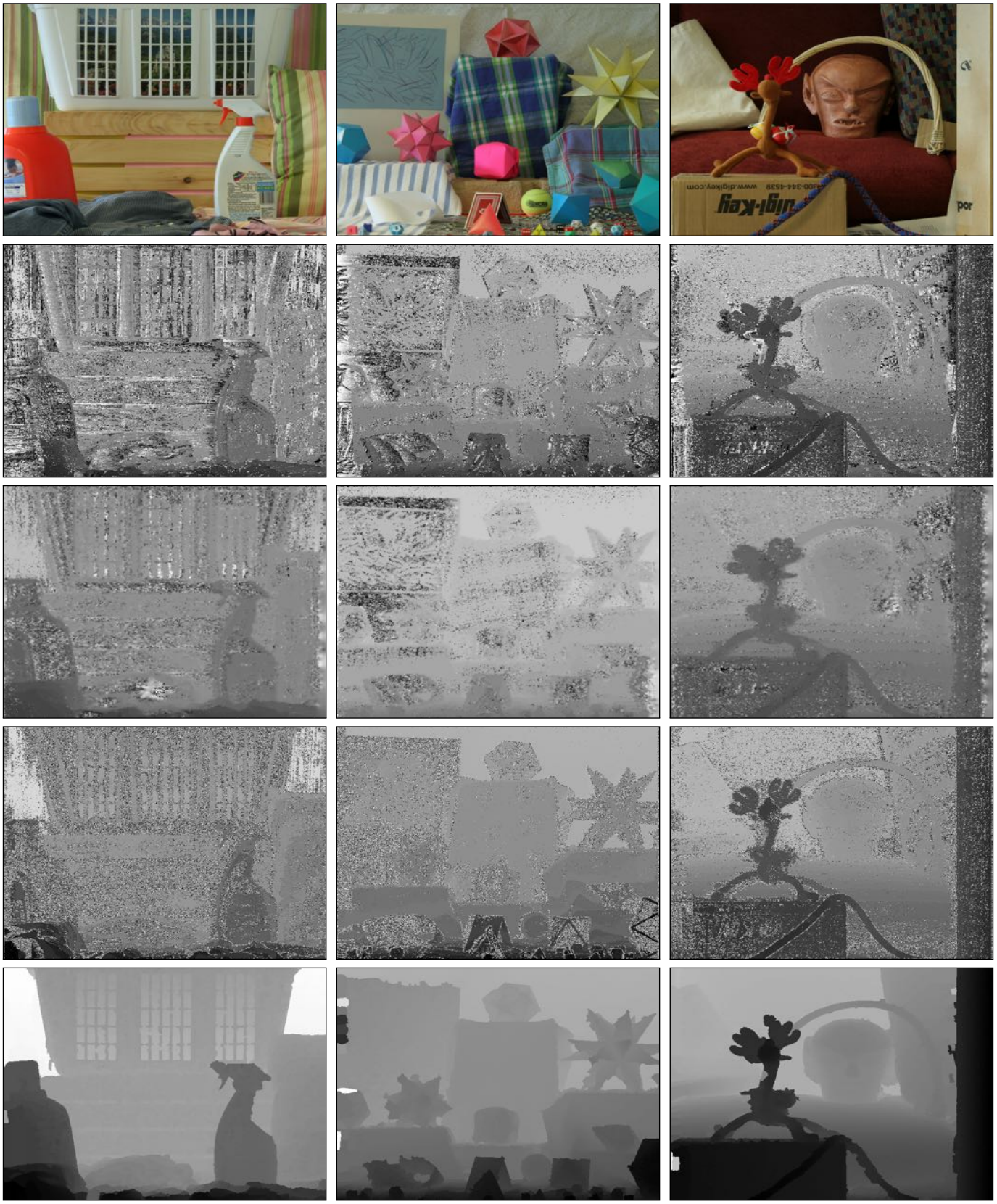

Figure 6: From top to bottom: the reference frame, the **Best Map**, the result of **MKS**, the result of **Kalman T** (temporal–only) and the result of **Kalman ST**. Images are automatically scaled in the range [0,255], hence the gray levels changes from row to row. Full resolution images can be seen on line [1].

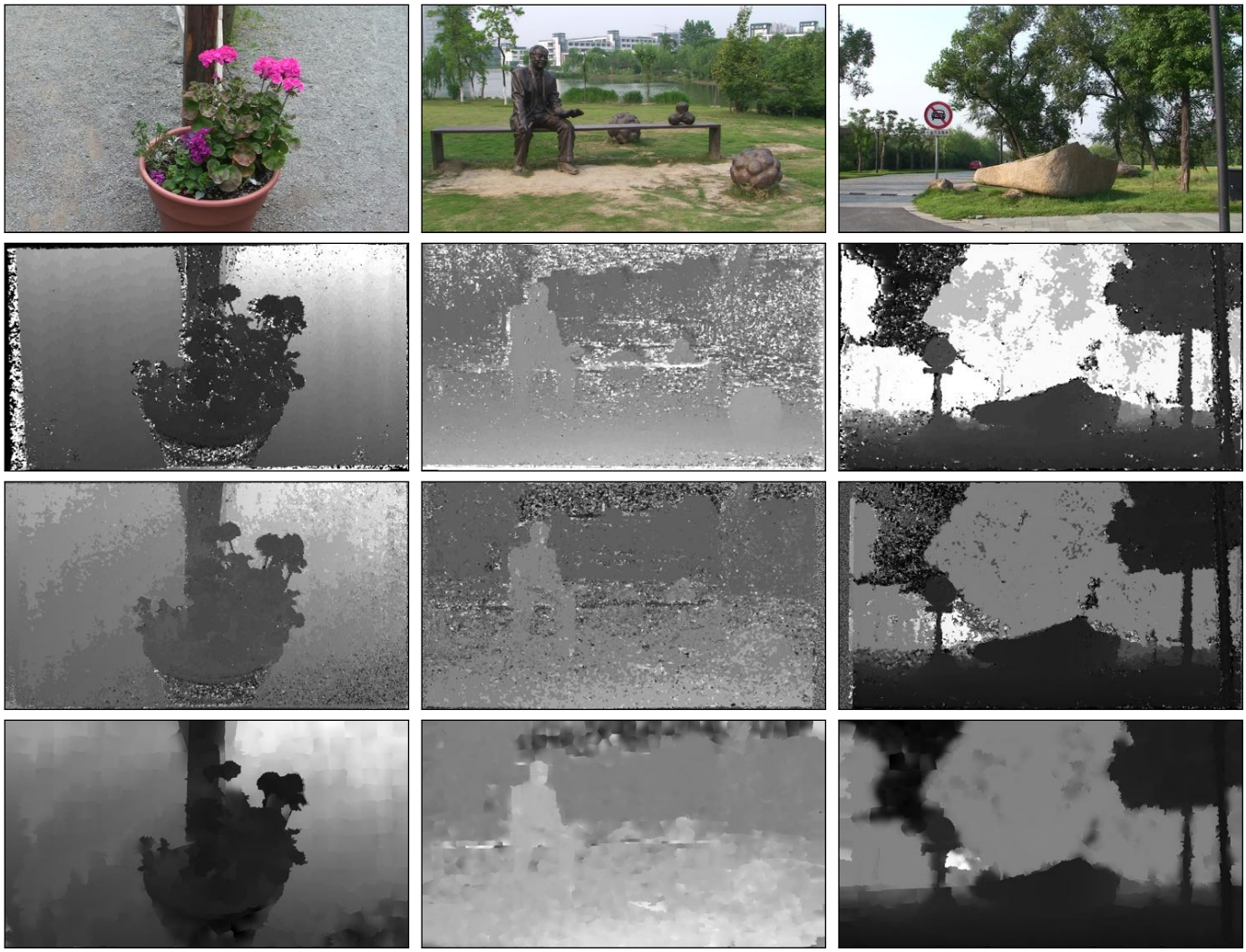

Figure 7: From top to bottom: the reference frame, the qualitative **Best Map** (manually selected), the result of **MKS**, the result of **Kalman T** (temporal–only) and the result of **Kalman ST**. Images are automatically scaled in the range [0,255], hence the gray levels changes from row to row. Full resolution images can be seen at www.diegm.uniud.it/fusiello/demo/dsp.

a depth-proxy and went trough all the stages required to compute it.

Since ground truth is not available, the evaluation will be only qualitative. Results, in Fig. 7, show a significant improvement on the strategy without spatial support, and are more consistent with the scene content, especially on occluded or badly measured pixels. MKS could not be evaluated on these images, as it is restricted to pure lateral motion.

5. Conclusion

In this paper we presented a framework that allows to combine parallax measurements obtained by processing the frames of a monocular video sequence. The integration takes place at two levels: i) temporal, where different estimates of depth values are merged along a timeline, and ii) spatial, where estimates are relaxed over pixel neighborhood. A segmentation into superpixels provides a spatial support that – in principle – does not cross objects boundaries.

Both spatial and temporal integration are derived as simple Kalman filters and are consistent with a data fusion framework based on the Mahalanobis distance [20]. They exploit confidence values provided by the stereo matching step. In our experiments all the confidence measures provided comparable results, so there is no clear indication that one measure is superior to the others. Instead, it turns out that singling out occlusions (with LRC) makes a real difference. The spatio–temporal integration has shown to be effective, and the benefits of the spatial step have been demonstrated with respect to the temporal-only version. The method is also compared with **MKS** and obtains consistently better results.

Appendix A. Confidence Measures

This appendix summarizes the different confidence measures that have been considered in the paper. The reader is referred to [8] for a more detailed description.

In the following $c(d)$ denotes the matching cost – normalized in $[0, 1]$ – associated to disparity hypothesis d. Since NCC is a similarity measure and all the confidence measures are defined using a cost function, $1 - \text{NCC}$ will be used instead.

The minimum cost for a pixel and its correspondent disparity value are respectively denoted by c_1 and d_1 (i.e. $c(d_1) = c_1 = \min(c(d))$). The second smallest cost value is denoted by c_2, while the second smallest value that is also a local minimum is represented by $c_{2\text{m}}$ (see Figure A.8).

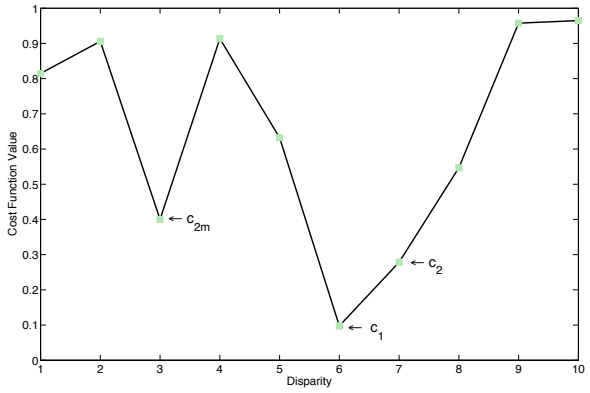

Figure A.8: Example of a cost function within a 10 pixel disparity range.

The confidence $\phi(i)$ varies in $[0, 1]$, where 0 means that the value at pixel i is *totally unreliable* and 1 means *fully confident*.

A very simple confidence metric is the matching score:

Matching Score (MSM):

$$\phi_{\text{MSM}} = 1 - c_1. \tag{A.1}$$

The first group of measures assess the cost function around its minimum by comparing it to the following smaller cost values (c_2 or $c_{2\text{m}}$) or to the disparity neighbors.

Curvature of the cost function (CUR):

$$\phi_{\text{CUR}} = \frac{2 + (-2c_1 + c(d_1 - 1) + c(d_1 + 1))}{4} \tag{A.2}$$

Peak Ratio (PKR):

$$\phi_{\text{PKR}} = 1 - \frac{c_1}{c_{2\text{m}}} \tag{A.3}$$

Maximum Margin (MMN):

$$\phi_{\text{MMN}} = \frac{c_2 - c_1}{c_2} \tag{A.4}$$

Winner Margin (WMN):

$$\phi_{\text{WMN}} = \frac{c_{2\text{m}} - c_1}{c_{2\text{m}}} \tag{A.5}$$

The following metrics take into account the entire cost curve and by assuming that it follows a normal distribution.

Maximum Likelihood Measure (MLM):

$$\phi_{\text{MLM}} = \frac{e^{-\frac{c_1}{2\sigma^2_{\text{MLM}}}}}{\sum_d e^{-\frac{c(d)}{2\sigma^2_{\text{MLM}}}}} \tag{A.6}$$

Attainable Maximum Likelihood (AML):

$$\phi_{\text{AML}} = \frac{1}{\sum_d e^{-\frac{(c(d)-c_1)^2}{2\sigma^2_{\text{AMLM}}}}} \tag{A.7}$$

We also considered two special measures to use as a touchstone. GT assigns confidence 1 if the corresponding pixel's disparity is correctly computed and 0 otherwise, according to the ground truth. UNI is an uninformed metric that assigns the same confidence to all the pixels.

Ground truth (GT):

$$\phi_{\text{GT}} = \begin{cases} 1 & \text{if disparity is correct} \\ 0 & \text{o/w} \end{cases} \tag{A.8}$$

Uniform (UNI):

$$\phi_{\text{UNI}} = \text{cost} \tag{A.9}$$

References

[1] www.diegm.uniud.it/fusiello/demo/dsp.

[2] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk. SLIC Superpixels Compared to State-of-the-art Superpixel Methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2274 – 2282, 2012.

[3] B. D. Anderson and J. B. Moore. *Optimal filtering*. Prentice-Hall information and system sciences series. Englewood Cliffs, N.J. Prentice-Hall, 1979.

[4] A. Fusiello and L. Irsara. Quasi-euclidean epipolar rectification of uncalibrated images. *Machine Vision and Applications*, 22(4):663 – 670, 2011.

[5] A. Fusiello, E. Trucco, and A. Verri. A compact algorithm for rectification of stereo pairs. *Machine Vision and Applications*, 12(1):16–22, 2000.

[6] M. Goesele, B. Curless, and S. M. Seitz. Multi-view stereo revisited. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 2402–2409. IEEE, 2006.

[7] C. Hernández and G. Vogiatzis. Shape from photographs: A multi-view stereo pipeline. In *Computer Vision: Detection, Recognition and Reconstruction*, volume 285 of *Studies in Computational Intelligence*, pages 281–311. Springer, Berlin, 2010.

[8] X. Hu and P. Mordohai. A quantitative evaluation of confidence measures for stereo vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2121–2133, 2012.

[9] M. Irani and P. Anandan. Parallax geometry of pairs of points for 3D scene analysis. In *Proceedings of the European Conference on Computer Vision*, pages 17–30, 1996.

[10] K. Kanatani. *Geometric Computation for Machine Vision*. Oxford University Press, Inc., New York, NY, USA, 1993.

[11] S. B. Kang, J. A. Webb, C. L. Zitnick, and T. Kanade. A multibaseline stereo system with active illumination and real-time image acquisition. In *Proceedings of the Fifth International Conference on Computer Vision*, pages 88–, 1995.

[12] M. Lang, O. Wang, T. Aydin, A. Smolic, and M. Gross. Practical temporal consistency for image-based graphics applications. *ACM Trans. Graph.*, 31(4):34:1–34:8, July 2012.

[13] H. Li and R. Hartley. Rectification-free multibaseline stereo for non-ideal configurations. In *Proceedings of the 13th international conference on Image Analysis and Processing*, pages 810–817, 2005.

[14] F. Malapelle, A. Fusiello, B. Rossi, E. Piccinelli, and P. Fragneto. Uncalibrated dynamic stereo using parallax. In *Proceedings of the 8th International Symposium on Image and Signal Processing and Analysis (ISPA)*, Trieste, Italy, 2013. IEEE.

[15] D. Marr and T. Poggio. Cooperative computation of stereo disparity. *Science*, 194(4262):283–287, 1976.

[16] L. Matthies, T. Kanade, and R. Szelisky. Kalman filter based algorithms for estimating depth from image sequences. *International Journal of Computer Vision*, 3:209–236, 1989.

[17] L. McMillan and G. Bishop. Head-tracked stereo display using image warping. In *Stereoscopic Displays and Virtual Reality Systems II*, number 2409 in SPIE Proceedings, pages 21–30, San Jose, CA, 1995.

[18] R. A. Newcombe and J. Andrew. Live dense reconstruction with a single moving camera davison. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

[19] M. Okutomi and T. Kanade. A multiple-baseline stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(4):353–363, 1993.

[20] R. Ramparany. An integrated support for fusion perceptual information. In Groen, Hirose, and Thorpe, editors, *Intelligent Autonomuos Systems*, pages 500–508. IOS Press, 1982.

[21] P. J. Rousseeuw and A. M. Leroy. *Robust regression & outlier detection*. John Wiley & sons, 1987.

[22] J. Santos-Victor and J. Sentiero. Generation of 3D dense depth maps by dynamic vision. In *British Machine Vision Conference*, pages 129–138, 1992.

[23] H. Sawhney. 3d geometry from planar parallax. In *Computer Vision and Pattern Recognition, 1994. Proceedings CVPR '94., 1994 IEEE Computer Society Conference on*, pages 929–934, Jun 1994.

[24] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, 47(1):7–42, 2002.

[25] S. Schwarz, M. Sjostrom, and R. Olsson. A weighted optimization approach to time-of-flight sensor fusion. *Image Processing, IEEE Transactions on*, 23(1):214–225, 2014.

[26] A. Shashua and N. Navab. Relative affine structure: Canonical model for 3D from 2D geometry and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(9):873–883, 1996.

[27] S. L. Spehn. Noise adaptation and correlated maneuver gating of an extended kalman filter. Naval Postgraduate School Monterey, CA, 1990.

[28] E. Trucco, V. Roberto, S. Tinonin, and M. Corbatto. SSD disparity estimation for dynamic stereo. In *Proceedings of the British Machine Vision Conference*, pages 342–352, 1996.

[29] J. K. Uhlmann. Covariance consistency methods for fault-tolerant distributed data fusion. *Information Fusion*, 4(3):201 – 215, 2003.

[30] G. Vogiatzis and C. Hernàndez. Video-based, real-time multi-view stereo. *Image and Vision Computing*, pages 434–441, 2011.

[31] A. M. Waxman and S. S. Sinha. Dynamic stereo: Passive ranging to moving objects from relative image flows. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(4):406–412, 1986.

[32] G. Zhang, J. Jia, T.-T. Wong, and H. Bao. Consistent depth maps recovery from a video sequence. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(6):974–988, 2009.