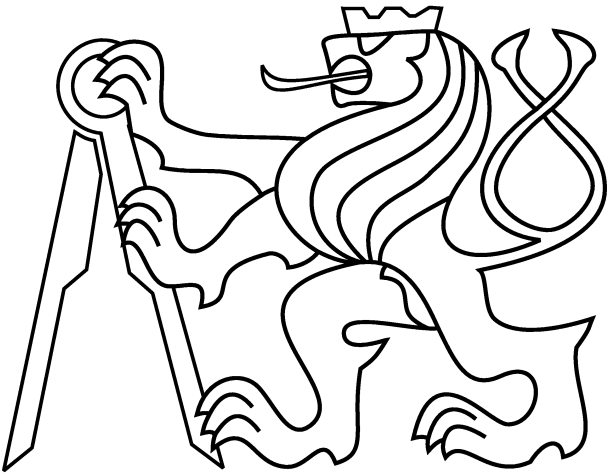
CZECH TECHNICAL UNIVERSITY IN PRAGUE



DOCTORAL THESIS STATEMENT

Czech Technical University in Prague
Faculty of Electrical Engineering
Department of Cybernetics

Jan Čech

Accurate and Robust Stereoscopic Matching in Efficient Algorithms

PhD Study Programme No. P 2612—Electrotechnics and Informatics,
branch No. 3902V035—Artificial Intelligence and Biocybernetics

Doctoral thesis statement for obtaining the academic title of "Doctor", abbreviated to "PhD".

Prague, February 2009

The doctoral thesis was produced in full-time manner PhD study at the department of Cybernetics of the Faculty of Electrical Engineering of the CTU in Prague

Candidate: Ing. Jan Čech
Department of Cybernetics,
Faculty of Electrical Engineering of CTU.

Thesis Advisor: Doc. Dr. Techn. Ing. Radim Šára
Department of Cybernetics,
Faculty of Electrical Engineering of CTU.

Prof. Ing. Vladimír Mařík, DrSc.
Chairman of the Board for the Defence of the Doctoral Thesis
in the branch of study No. 3902V035—
—Artificial Intelligence and Biocybernetics,
Department of Cybernetics, Karlovo náměstí 13, Prague 2

Abstract

The thesis studies dense stereoscopic techniques which are usable for accurate, robust and fast matching of high-resolution images of complex 3D scenes.

The main contributions are: (1) Image sampling invariant and affine insensitive complex correlation statistic (CCS) which is based on representing the image point neighbourhood as a response to complex Gabor filters. The CCS is a complex number with a magnitude of invariant similarity and a phase of estimated maximum position between pixels. (2) Methods for refining a disparity to sub-pixel precision - as an outcome of CCS phase, and alternatively also as a single continuous optimization problem based on a simple quadratic criterion. (3) A fast matching algorithm which avoids computing correlations for the entire disparity space by growing promising correspondence hypotheses from initial (even random) seeds. The growth is coupled with a confidently stable matching algorithm by Šára, ECCV 2002, which robustly selects the matching among competing hypotheses. (4) An algorithm for verification of given correspondences by uncalibrated dense matching. It is destined for selecting correspondences before RANSAC in challenging matching problems, with low ratio of inliers, in cluttered scenes where standard descriptor-based approach fails. An efficient procedure driven by Wald's sequential decision process grows a given correspondence while collecting statics until the decision based on learned models.

Some methods presented in the thesis go beyond the scope of 3D reconstruction, and they are applicable in many problems where the correspondences between images are sought.

Contents

1 Introduction

This thesis deals with various correspondence problems by dense (pixel-to-pixel) matching techniques. Points in the images are said to be *corresponding* if they are projections of an identical point in three-dimensional (3D) scene. In reality, we always search for a correspondence among pixels or larger image regions, although the correspondence is elementarily defined among image points.

The correspondence problem is one of the fundamental problems in computer vision. The tasks like 3D scene reconstruction from images, image registration, stitching, object recognition, image retrieval rely on correspondences among images. Our primary motivation is the automatic 3D reconstruction from images.

1.1 From unorganized images to a 3D model

We briefly introduce an automatic 3D reconstruction pipeline which was developed in Center for Machine Perception [7]. The algorithm produces a 3D model from a set of unorganized images.

A flowchart of our 3D reconstruction pipeline is sketched in Fig. 1. The algorithms works with unorganized images, which are captured by a consumer camera. An example is shown in Fig. 2. Notice, the object is occluded by people walking around and climbing the statue. This occurs in many images. Nevertheless, the algorithm is robust to this type of occlusions too.

First, distinguished regions are detected in all n input images. These are the regions which are repeatably detectable under image transformations induced by

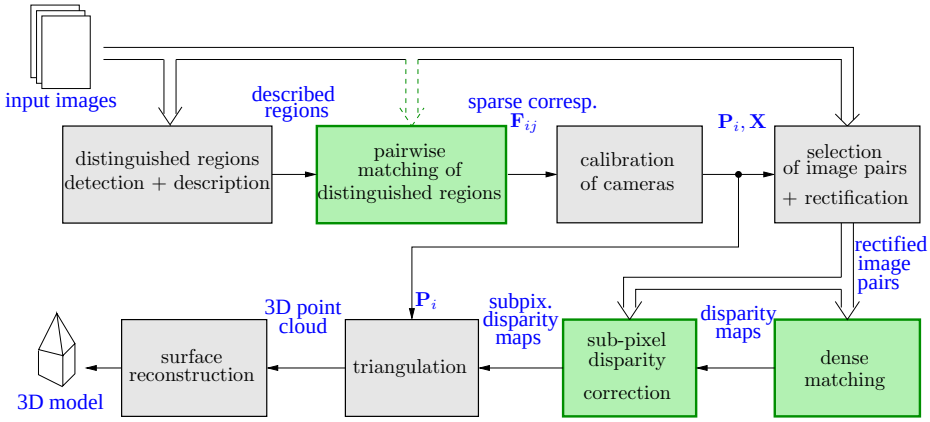


Fig. 1 3D reconstruction data processing pipeline. Blocks where the thesis contributes are highlighted by a green color.

Fig. 2 Input images of Henri Miller's statue L'écoute in Paris. The entire set used for 3D reconstruction contains 26 images.

various viewpoints. We used Maximally Stable Extremal Regions (MSERs) [23]. These regions are geometrically normalized according to regions's local affine coordinates constructed from local characteristics of the region. The normalized regions are then described with SIFT image descriptor [21]. The descriptor carries the viewpoint invariant (modelled as affine transformation invariant) information about the region's neighbourhood in the image.

Next, under no assumptions about the acquisition of images, each of $\binom{n}{2}$ image pairs are attempted to match in order to establish their epipolar geometry. Loosely speaking, the pairs of regions with closest descriptors are inserted into a set of tentative correspondences for each image pair [23]. The epipolar geometry is found by fitting the model of fundamental matrix $\mathbf{F}$ into the set of tentative correspondences using RANSAC. However, before the RANSAC is applied, we propose to verify the tentative correspondences first using an efficient dense matching-based algorithm which tries to match image neighbourhoods of tentative correspondences and evaluates their quality. This is important in difficult matching problems, where compact descriptors do not perform well, problems with high ratio (above 90 percent) of outliers among tentative correspondences, typically complex 3D structure with many occlusions. This method is described in Chapter 6, originally published in [4].

The correspondences between pairs of images which are inliers of the epipolar geometries are used for calibration of cameras. This is done by an algorithm [22]. The procedure robustly estimates the camera matrices $\mathbf{P}_i$ and sparse 3D points $\mathbf{X}$. The estimate of camera parameters and 3D points is refined by bundle-adjustment [34], which minimizes the reprojection error, i.e. the least squares error between the points detected in the images and reprojection of reconstructed points by estimated cameras.

Next, suitable image pairs for subsequent rectification and dense matching are selected. These are the image pairs with large number of sparse correspondences. The selected pairs are epipolarly rectified, i.e. the images are warped by homographies such that the epipolar lines are aligned with common image rows [24]. The rectification is important since it makes the distortion of local image neighbourhoods due to slanted surfaces well defined and allows the subsequent dense

(a) (b)

Fig. 3 The rectified stereo-pair with epipolar lines shown in a blue color (a). The color-coded disparity map; warmer colors correspond to higher disparities, colder color to lower disparities, gray color means unassigned disparity (b).

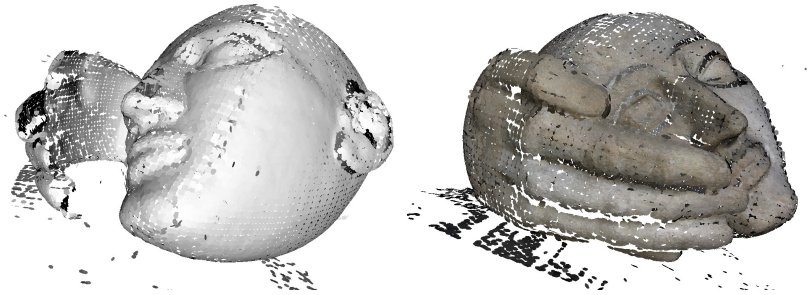


Fig. 4 Resulting 3D model as fish-scales [30]. The first view is untextured while the other is textured.

matching algorithm to be computationally efficient. The rectified stereopair is shown in Fig. 3(a).

The dense matching algorithm finds correspondences between pixels. For each rectified image pair, it produces a disparity map, which is inversely proportional to depth map. We use the algorithm described in Chapter 5 which was originally published in [6]. The disparity map for the above stereopair is shown in 3(b). Our approach is that the dense matching algorithm must not make too many errors, since the matching errors make the 3D model reconstruction difficult. However, the effort of making fewer matching errors is at the cost of a lower density of the disparity map. The matching algorithm has the 'reject' option, i.e. it does not decide a disparity in case of ambiguous data.

Integer disparity maps are refined to sub-pixel precision. Although we are using relatively high-resolution images (above 1 MPx), the sub-pixel disparity

correction visually improves the 3D model. This is possible by several approaches, our methods are described in Chapter 4, and partially published in [5].

Resulting sub-pixel disparity maps are 'unrectified', i.e. the correspondences are recalculated with respect to images before rectification, and triangulated using the estimated cameras. This results in a dense 3D point cloud.

The surface can be approximated with 'fish-scales' [30], small planar discs tangent to the surface. This simple method reduces the amount of data without significant loss of details and it is able to eliminate certain mismatches. The resulting 3D fish-scale model of the exemplar scene is shown in Fig. 4 in two different views both untextured and textured. If needed, the triangulated mesh is obtained by standard methods [1, 14].

2 State of the art

This section overviews state of the art in the dense stereoscopic matching, i.e. the algorithms which find correspondences between pixels as densely as possible up to the occlusions or ambiguity.

2.1 Taxonomy of dense matching methods

The outlined taxonomy is not strict and captures our view of the field.

There exist two basic approaches in dense matching methods: First, methods formulating the task as an explicit global optimization problem incorporating the prior model (A). Second, the methods where no explicit prior is employed and the correspondences are selected according to a certain principle (B).

2.1.1 Energy minimization methods (A)

These methods are also sometimes called *global* since the matching task is formulated as an optimization of a single criterion. The inherent ambiguity of the stereo matching is solved by regularization via incorporating a prior model of the scene.

The common justification of these methods is a probabilistic formulation in a Bayesian framework. The optimal disparity map $\mathbf{d}^*$ has the maximum aposteriori probability (MAP), i.e. the most probable solution given the input images $\mathbf{I_l}$, $\mathbf{I_r}$

$$\mathbf{d}^* = \arg\max_{\mathbf{d}\in\mathcal{D}} p(\mathbf{d} \mid \mathbf{I_l}, \mathbf{I_r}) = \arg\max_{\mathbf{d}\in\mathcal{D}} p(\mathbf{I_l}, \mathbf{I_r} \mid \mathbf{d}) p(\mathbf{d}). \tag{1}$$

We denote $\mathcal{D}$ a domain of disparity map $\mathbf{d}$. The first term $p(\mathbf{I_l}, \mathbf{I_r} \mid \mathbf{d})$ is referred as data term, it measures how well the current disparity map $\mathbf{d}$ maps the images $\mathbf{I_l}$ and $\mathbf{I_r}$ onto each other. The second term $p(\mathbf{d})$ is the prior term representing

the prior probability of the solution. Applying logarithm to the equation leads to energy minimization problem, where the product become a summation.

According to the domain $\mathcal{D}$ of the disparity map which is considered, we further distinguish the energy minimization methods. If domain $\mathcal{D}$ is a space of *discrete* functions of a discrete variable, we are speaking about discrete optimization methods (A.1). If domain $\mathcal{D}$ is a space of *continuous* (differentiable) functions, we are speaking about variational methods (A.2).

Discrete optimization methods (A.1) The disparity map is considered a finite set of pixels T with assigned discrete labels. We denote L a finite set of possible labels, which consists of a range of integer disparity values, and possibly some extra labels for occlusions.

The problem is often modeled using a Markov Random Field (MRF)

$$\mathbf{d}^* = \arg\min_{\mathbf{d}\in L^{|T|}} \sum_t \delta(\mathbf{I_l}, \mathbf{I_r}, d_t) + \lambda \sum_{t,t'} \rho(d_t, d_{t'}), \tag{2}$$

where $d_t \in L$ is a label which is assigned at pixel $t \in T$, δ is the unary potential, ρ is the binary potential. The summation is over all pixels t and over all pairs of neighbouring pixels t,t'. Weight λ controls the strength of the prior.

The above problem is NP-hard in general. There exist solvable sub-classes of the problem which are applicable to stereo [11, 31]. Solvability depends on the problem topology, structure of a label set and the form of pairwise potentials.

To avoid the NP-hard optimization, researchers decompose the 2D problem into a collection of 1D problems along epipolar lines, which are solved by dynamic programming, e.g. [13, 8], or select an acyclic sub-graph (a tree) of the underlying graph of the MRF, e.g. [35, 2]. Certain class of 2D problems can be solved by graph cuts algorithm, [3, 17], which are now considered successful. Other problems are optimized by approximate general solvers, e.g. belief propagation [33], TRW-S [15].

Variational methods (A.2) Following methods consider the domain of disparity map a space of continuous differentiable functions and express the problem (1) as a variational task

$$\mathbf{d}(x,y)^* = \arg\min_{\mathbf{d}(x,y)} \iint D\big(\mathbf{I_l}, \mathbf{I_r}, \mathbf{d}(x,y)\big) + \lambda R\big(\nabla \mathbf{d}(x,y)\big) dxdy, \tag{3}$$

where D is the data term, R the regularization (prior) term, x and y are the image coordinates. The Euler-Lagrange equation of the problem is a partial differential equation, which is solved by a gradient descent diffusion process [26, 32] or by level sets [9, 10].

2.1.2 Discriminability based correspondence selection (B)

These algorithms proceed such, that they recognize what belongs together based on the sufficient discriminability of the image point description without an explicit prior model of the scene. They use only simple geometrical constraints, like uniqueness and ordering.

Each image point is described by a feature vector. It is often a vector containing pixel intensities of a square neighbourhood of the image point (a window) or other more sophisticated descriptors. Then the algorithms construct a matching table, which is a set of all potential matches together with their similarity statistics computed over the feature vectors. The similarity statistic used is often SSD (Sum of Squared Differences), SAD (Sum of Absolute Differences), NCC (Normalized Cross Correlation), MNCC (Moravec's Normalized Cross Correlation) [25], etc.:

$$\mathrm{MNCC}(\mathbf{W}_i, \mathbf{W}_j) = \frac{2\,\mathrm{cov}(\mathbf{W}_i, \mathbf{W}_j)}{\mathrm{var}\,\mathbf{W}_i + \mathrm{var}\,\mathbf{W}_j}, \tag{4}$$

where the $\mathbf{W}_i$ and $\mathbf{W}_j$ are n-dimensional feature vectors of image point i and j respectively.

Finally, the matching algorithm establishes the matches based on the correlation table according to a certain *principle*. The simplest algorithm of this family is the Winner Takes All (WTA), which selects the match with the highest similarity for each row (or column) of the matching table, without any other constraints. Šára [29] revised the optimality condition and proposed a *stability* as a criterion for finding solution. The stability principle says that the solution should not change much with a small perturbation of data. He designed a Confidently Stable Matching (CSM) algorithm [28] which finds the largest unambiguous matching, according to a preselected confidence level. The algorithm constructs an oriented graph on the matching table cells in such a way that edges are oriented from the higher to lower similarity values within the forbidden zone generated by uniqueness and ordering constraint. Intervals around the similarity values derived from the confidence level are compared instead of the similarity values themselves. The task leads to finding the kernel of such graph.

2.2 Other Methods

There exist several other approaches in literature. The *progressive* methods establishes the matches iteratively. Early (more confident) matches reduce the ambiguity of the subsequent matches, e.g. [36]. The *phase-based* techniques are based on the Fourier shifting theorem: a shift in the spatial domain corresponds to the phase shift in the frequency domain, e.g. [27, 12]. The *space carving* methods solve the correspondence problem in a reversed way. The observed scene is

divided into voxels and the scene is reconstructed such that rays from each voxel to all cameras are projected to the images. Each voxel is either filled or empty based on the variance of the projection into the images, e.g. [20].

3 Program of the thesis

After reviewing the state-of-the-art methods and getting experience from the algorithm evaluation [18, 19], we decided to study and develop methods of class B. The methods of class A using strong prior models suffer from several types of artifacts: they smooth out occlusion boundaries and interpolate through low-textured areas, see Fig. 5 (a), (b). This is due to inability of their prior model to capture well the statistics of natural scenes. On the other hand, our conception is based on the sufficient discriminability of data. The proposed algorithm has the 'reject-option', such that the ambiguous regions are identified and the disparity is assigned in reliable regions only, see Fig. 5 (c), (d).

3.1 Problems of the standard approach

Discretization artifacts. A sampling of the images corrupts standard correlation statistics. This effect is the stronger the higher frequencies are present in the images.

No invariance to affine distortion. Using a standard similarity statistic, e.g. (4), computed from raw image intensities in square windows assumes the fronto-parallel planar setup. This is violated whenever cameras observe a slanted plane. The matching window is geometrically distorted, it becomes stretched or skewed.

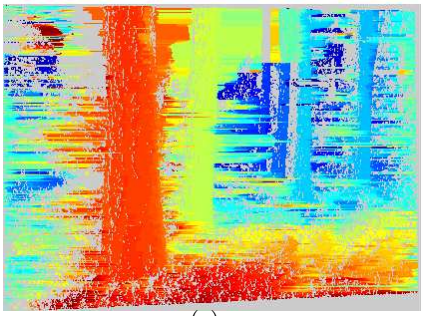
Too large number of similarity statistics to be computed. A standard approach is to compute the correlation statistics for all the potential matches, i.e. the entire disparity space. However, the disparity space is large for high-resolution images. Consider a pair of square 1 Mpx images, 1000×1000 pixels. Then one needs to compute all $1000 \times 1000 \times 1000 = 1$billion of correlation values, which takes time.

3.2 Contribution of the thesis

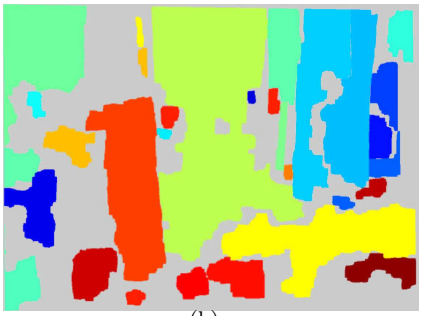
1. **Sampling invariant correlation statistic.** We designed the Complex Correlation Statistic, which is both invariant to image discretization and it also provides an estimate of the sub-pixel translation between the local image neighbourhoods. The value is not a real but a complex number. The magnitude is the invariant similarity, while the phase is the estimate of the sub-pixel shift. The main idea is that the pixel signatures are represented by

left rectified image

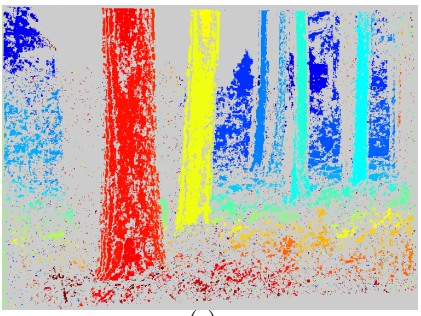
right rectified image



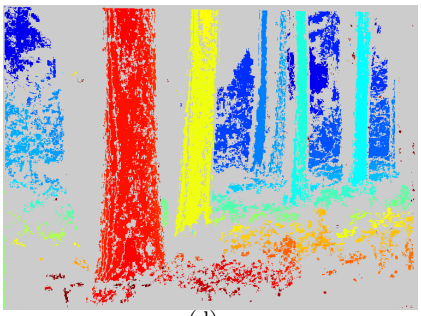
(a)



(b)



(c)



(d)

Fig. 5 Larch grove data set. Disparity maps from (a) Dynamic Programming [8], (b) Graph Cuts [16], (c) Confidently Stable Matching [28], (d) Growing Correspondence Seeds [6] (the proposed method). *Disparity maps (a) and (b) by courtesy of Jana Kostlivá.*

responses to a bank of complex Gabor filters. The statistic is then computed from the responses in a closed form. This work has been published in [5].

2. **Affine insensitive correlation statistic.** In a second step, we incorporated insensitivity to affine distortion of a local image neighbourhood, which occurs due to a surface slant, into the Complex Correlation Statistic. The basic idea is to decompose the distortion into orthogonal aspects (stretch and skew) and design a filter bank which is able to compensate an affine distortion in a limited range.

3. **Global method for sub-pixel disparity correction.** We proposed a method of a sub-pixel disparity correction which is formulated as a single optimization problem for the entire image. It is a continuous optimization problem based on a simple quadratic criterion. The algorithm finds a sub-pixel disparity map directly by a gradient descent from a given initialization by an integer disparity map.

4. **Fast stereo matching algorithm which involves computation of a small fraction of disparity space.** It turns out, it is not necessary to compute the correlations exhaustively for the entire disparity space. We propose an algorithm which visits only a small fraction of the disparity space (less than 1 per cent) while it still keeps a good performance, in the sense of robustness to occlusions, weak textures and repetitive patterns, as an exhaustive algorithm. Hereby, the speed up achieved is in the order of two magnitudes. The main idea is to generate promising correspondence hypotheses by growing initial correspondences (seeds). Finally, a robust matching algorithm is applied to select the stable matching [28] among the competing correspondence hypotheses. The algorithm is not dependent on given high-quality seeds, but surprisingly it works with seeds which are generated completely randomly. This work has been published in [6].

5. **Verification of tentative correspondences based on fast dense--matching algorithm.** We apply the growing mechanism to construct a procedure which grows a given tentative correspondence, collecting primitive statistics, in order to estimate how likely it is a true correspondence or a mismatch. Such a verification is typically used before fitting a model using RANSAC. It has a large impact in difficult matching problems where the ratio of outliers is above 90 per cent, where matching of standard compact descriptors fails usually due to a complex 3D structure with many occlusions. The procedure is driven by Wald's sequential decision which makes the verification process efficient. The time spent by the verification is negligible with respect to the time spent by matching of tentative correspondences. The method is originally published in [4].

4 Complex Correlation Statistic and sub-pixel disparity

A traditional solution of area-based stereo uses some kind of windowed pixel intensity correlation. This approach suffers from discretization artifacts which corrupt the correlation value. We introduce a new correlation statistic, which is completely invariant to image sampling, moreover it naturally provides a position of the correlation maximum between pixels. Additionally, we present a version which is insensitive to affine distortion of corresponding neighbourhoods which occurs due to a surface slant. Hereby we can obtain sub-pixel disparity directly from invariant and highly discriminable measurements without any postprocessing of the integer disparity map.

The key idea behind is to represent the image point neighbourhood as a response to a bank of Gabor filters. The images are convolved with the filter bank and the complex correlation statistic (CCS) is evaluated from the responses without iterations. The magnitude of CCS measures the image similarity and the phase gives the sub-pixel position.

We also present a simple global method for sub-pixel disparity correction which is posed as a single optimization task and solved iteratively, a complementary approach to sub-pixel disparities from CCS. The criterion consists of a quadratic data term penalizing discrepancies between the target image and the reference image warped according to estimated disparity map, and a smoothness term which penalizes differences in neighbouring disparities also quadratically.

The CCS statistic is computed for all potential integer matches between the rectified images $\mathbf{I_l}(x_1, y)$ and $\mathbf{I_r}(x_2, y)$, such that the images are represented as responses to a bank of complex Gabor filters tuned to different frequencies

$$c_i(x, y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} e^{j(u_{0i}x + v_{0i}y)}. \tag{5}$$

The responses are computed as convolutions

$$G_i^l(x, y) = \mathbf{I_l} * c_i, \quad G_i^r(x, y) = \mathbf{I_r} * c_i, \quad G_{xi}^l(x, y) = \mathbf{I_l} * \frac{\partial}{\partial x} c_i. \tag{6}$$

Each tuning frequency is used to estimate the local frequency u_i^l and local subpixel disprity Δ_i

$$u_i^l(x, y) = \frac{\Im(G_{xi}^l)\Re(G_i^l) - \Re(G_{xi}^l)\Im(G_i^l)}{|G_i^l|^2}, \tag{7}$$

$$\Delta_i(x_1, x_2) = \frac{\arg(\overline{G_i^l(x_1, y)}\, G_i^r(x_2, y))}{u_i^l(x_1, y)}. \tag{8}$$

These estimates are finally aggregated by CCS formula

$$\mathrm{CCS}(x_1,x_2) = \frac{2\sum_{i=1}^{N} |G_i^l(x_1,y)|\,|G_i^r(x_2,y)|\,e^{j\Delta_i(x_1,x_2)}}{\sum_{i=1}^{N} |G_i^l(x_1,y)|^2 + \sum_{i=1}^{N} |G_i^r(x_2,y)|^2} = A(x_1,x_2)\,e^{j\delta(x_1,x_2)}. \quad (9)$$

Magnitude A is the sampling invariant measurement, phase δ is the estimate of correlation maximum between pixels. The affine version $\mathrm{CCS}_{\mathrm{aff}}$ has magnitude A additionally insensitive to a small affine distortion of the matching window which occurs due to a surface slant. This is achieved by a decomposition of the distortion into orthogonal aspects (stretch, skew) and by a special form of the filter bank.

The other approach, the global method for the sub-pixel disparity correction, stands on the direct minimization of the functional

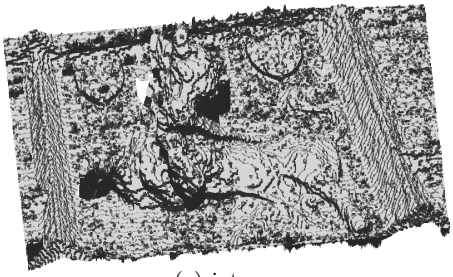
$$\begin{aligned} J(\mathbf{d}) = &\sum_{x=1}^{N}\sum_{y=1}^{M}\big(\mathbf{I_l}(x,y) - \mathbf{I_r}(x+\mathbf{d}_{x,y},y)\big)^2 + \\ &+ \lambda \sum_{x=1}^{N-1}\sum_{y=1}^{M-1}\big((\mathbf{d}_{x,y}-\mathbf{d}_{x+1,y})^2 + (\mathbf{d}_{x,y}-\mathbf{d}_{x,y+1})^2\big). \end{aligned} \quad (10)$$

for the disparity map $\mathbf{d}(x,y) = \mathbf{d}_{x,y}$. The quasi-Newton minimization with analytical gradient is started from the integer disparity map with marked occlusions. The quadratic regularization is 'disconnected' over occlusion boundaries.

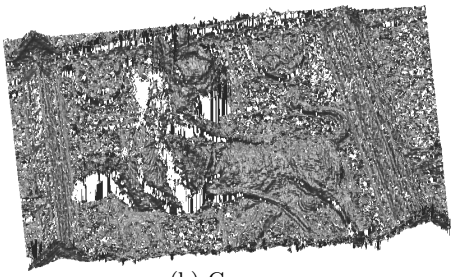
Results of sub-pixel correction methods are shown in Fig. 6. All the methods were initialized by the same integer disparity map (a) obtained by the proposed GCS-2 algorithm. Other reference methods are $\mathrm{C}_{\mathrm{interp}}$ which is based on parabola fitting into 3 points of MNCC statistic, and `dispcor` which is based on fitting an affine window.

5 Efficient sampling of disparity space

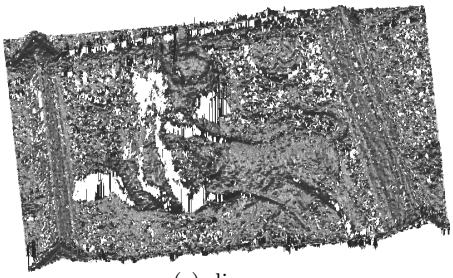
A simple stereo matching algorithm is proposed that visits only a small fraction of disparity space in order to find a semi-dense disparity map. It works by growing from a small set of correspondence seeds. Unlike in known seed-growing algorithms, it guarantees matching accuracy and correctness, even in the presence of repetitive patterns. The proposed algorithm is able to work with complex scenes with a rich 3D structure of a great depth and many complicated occlusions, see Fig. 7. This success is based on the fact it solves a kind of a global optimization



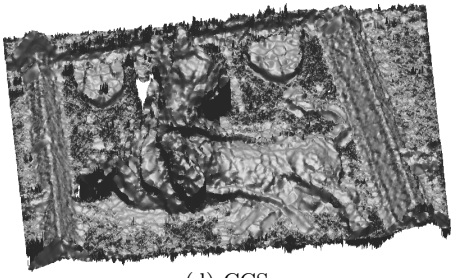
(a) integer



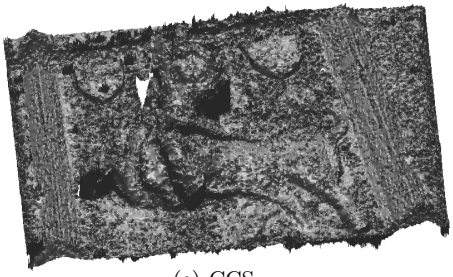
(b) C_{interp}



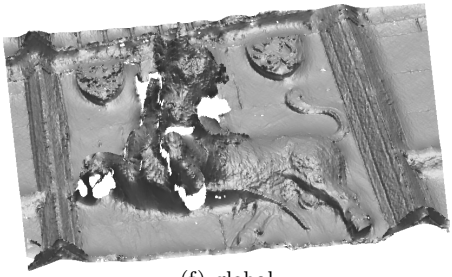
(c) dispcor



(d) CCS



(e) CCS_{aff}



(f) global

Fig. 6 Bull. Input images and reconstruction results as relighted 3D models.

Fig. 7 Complex scenes and matching results as disparity maps obtained by the proposed GCS-2 algorithm.

task. The algorithm can recover from wrong initial seeds to the extent they can even be random. The quality of correspondence seeds influences computing time, not the quality of the final disparity map.

The proposed algorithm avoids computing correlations for all potential matches in the entire disparity space by growing high similarity disparity components. Given an initial correspondence seed, the growing procedure searches for another correspondence in a small neighbourhood of the seed based on correlation. If a new correspondence is found, it becomes a new seed and the process repeats until all the seeds are exhausted. The final decision is performed by Confidently Stable Matching [28] which selects among competing correspondence hypotheses. The algorithm is *not* a direct combination of unconstrained growth followed by a filtration step. Instead, the growing procedure is designed to fit the final matching algorithm in the sense the growing is stopped whenever a correspondence hypothesis cannot win the final matching competition.

We show that the proposed algorithm achieves similar results as an exhaustive disparity space search but it is two orders of magnitude faster. This is very unlike the existing growing algorithms which are fast but erroneous. Accurate matching

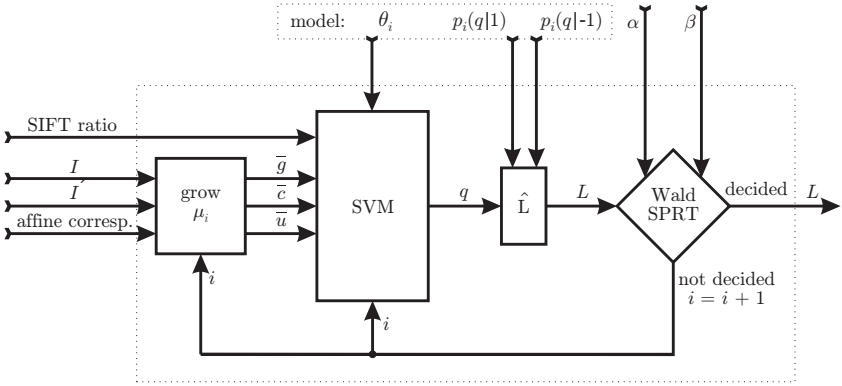


Fig. 8 The Sequential Correspondence Verification (SCV) algorithm.

on 2-megapixel images of complex scenes is routinely obtained in a few seconds on a common PC from a small number of seeds, without limiting the disparity search range.

6 Efficient sequential correspondence selection by cosegmentation

In many retrieval, object recognition and wide baseline stereo methods, correspondences of interest points (distinguished regions, transformation covariant points) are established possibly sublinearly by matching a compact descriptor such as SIFT. We show that a subsequent cosegmentation process coupled with a quasi-optimal sequential decision process leads to a correspondence verification procedure that has (i) high precision (is highly discriminative) (ii) good recall and (iii) is fast.

A flowchart of the algorithm is shown in Fig. 8. Given the input images $\mathbf{I}$, $\mathbf{I}'$, an affine correspondence to be verified and its ratio of the first to second closest SIFT descriptors, the algorithm first tries to perform a limited number of growing steps μ. It collects the primitive statistics from the cosegmented region $(\bar{g}, \bar{c}, \bar{u})$ which are combined with SIFT ratio by SVM to produce a scalar statistical correspondence quality q. This projection is in order to avoid estimating 4-dimensional probability distributions. Likelihood ratio $L = p_i(q \mid +1)/p_i(q \mid -1)$ is computed. Wald's SPRT test with given false-positive, false-negative thresholds α, β is performed. If the test is conclusive, the final likelihood ratio L is assigned besides the decision as an output. Otherwise, another decision stage with a larger number of growing steps μ is performed, i.e. with potentially larger cosegmented regions.

This process repeats until the decision or until the maximum number of growing steps μ is reached.

Models, SVM weights θ_i and conditional probability distributions $p_i(q \mid +1)$ and $p_i(q \mid -1)$, were learned from a set of 16000 exemplar correspondences collected from easy wide-baseline-stereo problems.

Experimentally we show that the process significantly outperforms the standard correspondence selection process based on SIFT distance ratios on challenging matching problems. A challenging example is shown in Fig. 9, which shows an image pair, where one image is captured in winter while the other in summer. The proposed SCV algorithm discovered several correct correspondences. Note that the results are shown before RANSAC validation by the epipolar geometry model. Blue parallelograms are the measurement regions of the SIFT descriptors, i.e. image regions where the SIFT descriptors are constructed from. Notice, they overreach the surfaces and make the descriptors contaminated with a background clutter. On the other hand, the SCV algorithm collects the statistics from correctly cosegmented 3D surfaces shown in a red color. This is the fundamental advantage of the proposed algorithm compared to a standard approach.

7 Conclusion

The first part of the thesis was dealing with low-level processing of image signals. We proposed a complex correlation statistic which possesses the invariance to image discretization and insensitivity to affine transformation of corresponding domains. The complex correlation statistic naturally provides an estimate of its maximum position with sub-pixel precision.

The second part of thesis was more algorithmic. We designed an efficient dense matching algorithm suitable for matching high-resolution images of complex 3D scenes. It avoids visiting the entire disparity space and computing huge number of correlation statistics by a sampling strategy. Only promising correspondence hypotheses are generated via growing initial correspondence seeds, which can be even random. Final decision is performed by a robust matching of competing correspondence hypotheses. The proposed algorithm keeps the robust properties (low error rate, sufficient density) of the original algorithm which computes the disparity space exhaustively, however it runs about 100 times faster.

The last part of the thesis originated from an inspiration by the success of the growing principle. This time we employed a similar growing mechanism to a problem of correspondence verification. We designed a procedure which estimates how likely a given correspondence is correct and how likely it is a mismatch. The algorithm, driven by Wald's sequential decision process, successively expands potentially corresponding regions while collecting image statistics until the decision based on learned models. We showed the resulting correspondence

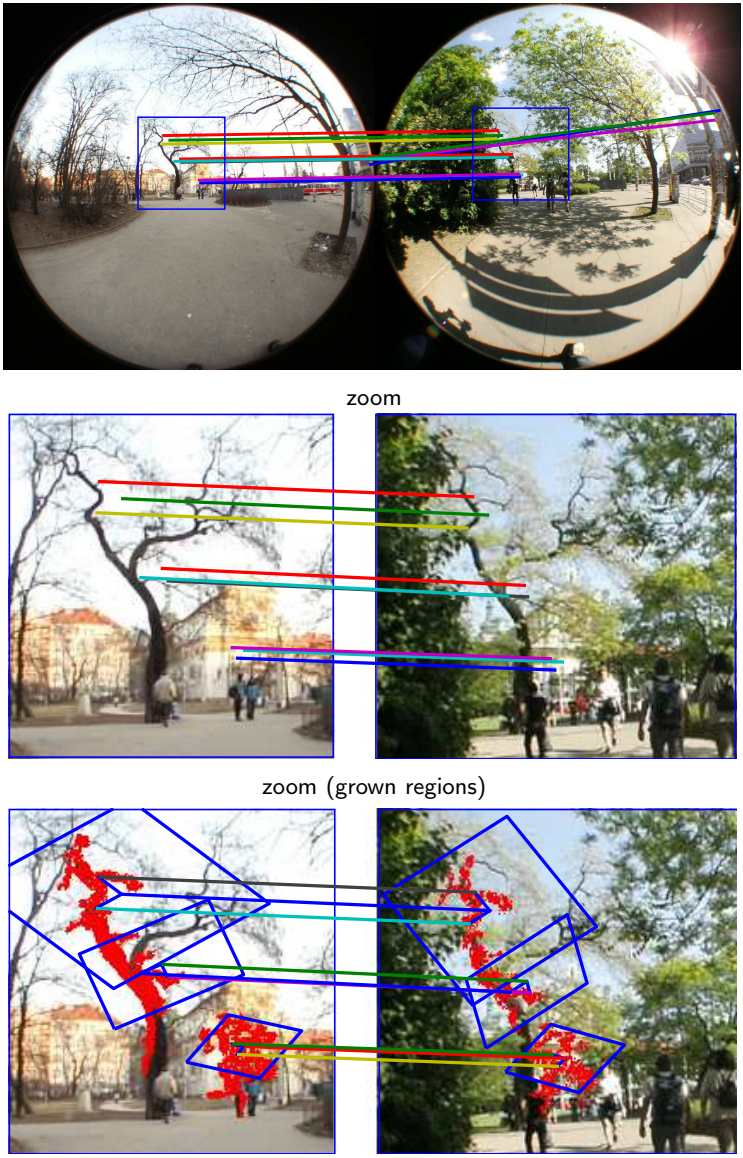


Fig. 9 Correspondences found by the SCV algorithm in challenging omnidirectional image pair with a 'wide temporal baseline'. *Raw images by courtesy of Jan Knopp.*

selection procedure is very discriminative, outperforms state-of-the-art correspondence selection based on ratio of two closest SIFT descriptors. Its running time is negligible to time spent by matching tentative correspondences, which implies that it should be always used before RANSAC. We showed benefits in challenging wide-baseline stereo problems and in image retrieval.

Contributions of the thesis are summarized in Sec. 3.2. A list of author's publications, their citations and other community acceptance can be found in Appendices.

References

[1] N. Amenta, M. Bern, and D. Eppstein. The crust and the β-skeleton: Combinatorial curve reconstruction. *Graphical Models and Image Processing*, 60(2):125–135, 1988.

[2] M. Bleyer and M. Gelautz. Simple but effective tree structures for dynamic programming-based stereo matching. In *Proc. Internation Conference on Computer Vision Theory and Applications (VISAPP)*, pages 415–422, 2008.

[3] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. In *Proc. ICCV*, volume 1, pages 377–384, 1999.

[4] J. Čech, J. Matas, and M. Perd'och. Efficient sequential correspondence selection by cosegmentation. In *Proc. CVPR*, 2008.

[5] J. Čech and R. Šára. Complex correlation statistic for dense stereoscopic matching. In *Proc. SCIA*, volume LNCS 3540, pages 598–608, 2005.

[6] J. Čech and R. Šára. Efficient sampling of disparity space for fast and accurate matching. In *Proc. CVPR Workshop Towards Benchmarking Automated Calibration, Orientation, and Surface Reconstruction from Images, BenCOS 2007*. IEEE Computer Society, 2007.

[7] H. Cornelius, R. Šára, D. Martinec, T. Pajdla, O. Chum, and J. Matas. Towards complete free-form reconstruction of complex 3D scenes from an unordered set of uncalibrated images. In *Proc. ECCV Workshop Statistical Methods in Video Processing*, LNCS 3247, pages 1–12, 2004.

[8] I. J. Cox, S. L. Hingorani, and S. B. Rao. A maximum likelihood stereo algorithm. *Computer Vision and Image Understanding*, 63(3):542–567, 1996.

[9] R. Deriche, C. Bouvin, and O. Faugeras. A level-set approach for stereo. In *First Annual Symposium on Enabling Technologies for Law Enforcement and Security - SPIE Conference 2942: Investigative Image Processing*, 1996.

[10] O. Faugeras and R. Keriven. Complete dense stereovision using level set methods. In *Proc. ECCV*, LNCS 1406, pages 379–393, 1998.

[11] B. Flach and M. I. Schlesinger. A class of solvable consistent labeling problems. In *Proc. IAPR Workshop on Advances in Pattern Recognition*, volume LNCS 1876, pages 652–658, 2000.

[12] D. J. Fleet, A. D. Jepson, and M. R. M. Jenkin. Phase-based disparity measurement. *CVGIP: Image Understanding*, 53(2):198–210, 1991.

[13] G. L. Gimel'farb, V. B. Marchenko, and V. I. Rybak. Algorithm of automatic matching of identical patches in stereopairs. *Kibernetika*, (2):118–129, 1972. In Russian.

[14] M. Kazhdan, M. Bolitho, and H. Hoppe. Poisson surface reconstruction. In *Proc. Eurographics Symposium on Geometry Processing*, pages 61–70, 2006.

[15] V. Kolmogorov. Convergent tree-reweighted message passing for energy minimization. *IEEE Trans. on PAMI*, 28(10):1568–1583, 2006.

[16] V. Kolmogorov and R. Zabih. Computing visual correspondence with occlusions using graph cuts. In *Proc. ICCV*, pages 508–515, 2001.

[17] V. Kolmogorov and R. Zabih. What energy functions can be minimized via graph cuts. *IEEE Trans. on PAMI*, 26(2):147–159, 2004.

[18] J. Kostková, J. Čech, and R. Šára. Dense stereomatching algorithm performance for view prediction and structure reconstruction. In *Proc. SCIA*, LNCS 2749, pages 101–107, 2003.

[19] J. Kostlivá, J. Čech, and R. Šára. Feasibility boundary in dense and semi-dense stereo matching. In *Proc. CVPR Workshop Towards Benchmarking Automated Calibration, Orientation, and Surface Reconstruction from Images, BenCOS 2007*. IEEE Computer Society, 2007. Best Paper Award.

[20] K. N. Kutulakos and S. M. Seitz. A theory of shape by space carving. In *Proc. ICCV*, volume 1, pages 307–314, 1999.

[21] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[22] D. Martinec. *Robust Multiview Reconstruction*. PhD thesis, Center for Machine Perception, K13133 FEE Czech Technical University, Prague, Czech Republic, 2008.

[23] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In *Proc. BMVC*, pages 384–393, 2002.

[24] M. Matoušek. *Epipolar Rectification Minimising Image Loss*. PhD thesis, Center for Machine Perception, K13133 FEE Czech Technical University, Prague, Czech Republic, 2007.

[25] H. P. Moravec. Towards automatic visual obstacle avoidance. In *Proc. International Joint Conference on Artificial Intelligence*, page 584, 1977.

[26] L. Robert and R. Deriche. Dense depth map reconstruction: A minimization and regularization approach which preserves discontinuities. In *Proc. ECCV*, LNCS 1064, pages 439–451, 1996.

[27] T. D. Sanger. Stereo disparity computation using Gabor filters. *Biological Cybernetics*, 58(6):405–418, 1988.

[28] R. Šára. Finding the largest unambiguous component of stereo matching. In *Proc. ECCV*, LNCS 2352, pages 900–914, 2002.

[29] R. Šára. Robust correspondence recognition for computer vision. In *COMPSTAT 2006 - Proceedings in Computational Statistics*, pages 119–131. Physica-Verlag, 2006.

[30] R. Šára and R. Bajcsy. Fish-scales: Representing fuzzy manifolds. In *Proc. ICCV*, pages 811–817, 1998.

[31] M. I. Schlesinger and B. Flach. Analysis of optimal labelling problems and their application to image segmentation and binocular stereovision. In *East-West-Vision: International Workshop & Project Festival on Computer Vision, Computer Graphics, New Media*, pages 55–60. Austrian Computer Society, 2002.

[32] C. Strecha, T. Tuytelaars, and L. Van Gool. Dense matching of multiple wide-baseline views. In *Proc. ICCV*, 2003.

[33] J. Sun, H.-Y. Shum, and N.-N. Zheng. Stereo matching using belief propagation. In *Proc. ECCV*, LNCS 2351, pages 510–524, 2002.

[34] B. Triggs, P. McLauchlan, R. Hartley, and A. Fitzgibbon. Bundle adjustment – a modern synthesis. In *Vision Algorithms: Theory and Practice*, LNCS 1883, pages 298–372, 1999.

[35] O. Veksler. Stereo correspondence by dynamic programming on a tree. In *Proc. CVPR*, volume 2, pages 384–390, 2005.

[36] Y. Wei and L. Quan. Region-based progressive stereo matching. In *Proc. CVPR*, pages 106–113, 2004.

A Resumé in Czech

Disertační práce studuje techniky hustého stereoskopického párování, které jsou použitelné pro přesné, robustní a rychlé párování obrazů ve vysokém rozlišení zachycujících složitou 3D scénu.

Hlavní příspěvky jsou: (1) Na vzorkování obrazu invariantní a affinně necitlivá komplexní korelační statistika (CCS), která je založena na reprezentaci okolí obrazového bodu pomocí odezev Gaborových filtrů. Statistika CCS je komplexní číslo, jehož absolutní hodnota představuje invariantní podobnost obrazových okolí, fáze představuje odhad polohy maxima mezi pixely. (2) Metody pro zpřesnění disparity na pod-pixelovou úroveň - použitím výsledeku z fáze CCS a alternativně jako jediný spojitý optimalizační problém, který je založený na jednoduchém kvadratickém kritériu. (3) Rychlý párovací algoritmus, který se vyhýbá vypočítávání korelací pro celý disparitní prostor tak, že narůstá slibné korespondenční hypotézy z počátečních (dokonce i náhodných) semínek. Narůstání je spojeno se stabilním párovacím algoritmem Šára, ECCV 2002, který nakonec robustně vybere párování mezi soutěžícími hypotézami. (4) Algoritmus pro ověření správnosti dané korespondence použitím nekalibrovaného hustého párování. Algoritmus je navržen pro výběr korespondencí před procedurou RANSAC v obtížných párovacích problémech, s nízkým poměrem korespondencí vyhovujících modelu, ve scénách se složitým rušivým pozadím, kde standardní přístup založený na desktriptorech selhává. Efektivní procedura řízená Waldovým rozhodovacím procesem narůstá danou korespondenci a sbírá statistiky až do rozhodnutí, které je založeno na naučených modelech.

Některé metody představené v disertační práci sahají nad rámec 3D rekonstrukce a jsou aplikovatelné v mnoha problémech počítačového vidění, kde jsou hledány korespondence.

B Author's publications

Journal papers:

[P1] J. Čech, J. Matas, and M. Perd'och. Efficient sequential correspondence selection by cosegmentation. *IEEE Trans. on PAMI*. In Review. Authorship 50-30-20.

[P2] J. Čech and R. Šára. Languages for constrained binary segmentation based on maximum aposteriori probability labeling. *International Journal of Imaging Systems and Technology*. To Appear. Authorship 50-50.

[P3] H.-F. Zhang, J. Čech, R. Šára, F.-C. Wu, and Z.-Y. Hu. Theory and robust algorithm of trinocular rectification. *Chinese Journal of Software*, 15(5):676–688, May 2004. Authorship 35-25-20-10-10.

Conference papers:

[P4] S. Aksoy, B. Özdemir, S. Eckert, F. Kayitakire, M. Pesaresi, O. Aytekin, C. C. Borel, J. Čech, E. Christophe, S. Düzgün, A. Erener, K. Ertugay, E. Hussain, J. Inglada, S. Lefèvre, O. Ok, D. K. San, R. Šára, J. Shan, J. Soman, I. Ulusoy, and R. Witz. Performance evaluation of building detection and digital surface model extraction algorithms: Outcomes of the PRRS 2008 algorithm performance contest. In *5th IAPR Workshop on Pattern Recognition in Remote Sensing*, Tampa, USA, December 2008. Authorship 22×100/22.

[P5] J. Čech, J. Matas, and M. Perd'och. Efficient sequential correspondence selection by cosegmentation. In *CVPR 2008: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Anchorage, USA, June 2008. Omnipress. Authorship 50-30-20.

[P6] J. Čech and R. Šára. Windowpane detection based on maximum aposteriori probability labeling. In R. P. Barneva and V. Brimkov, editors, *Image Analysis - From Theory to Applications, Proceedings of the 12th International Workshop on Combinatorial Image Analysis (IWCIA'08)*, pages 3–11, Buffalo, USA, April 2008. Research Publishing Services. Authorship 50-50.

[P7] J. Čech and R. Šára. Efficient sampling of disparity space for fast and accurate matching. In *BenCOS 2007: CVPR Workshop Towards Benchmarking Automated Calibration, Orientation and Surface Reconstruction from Images*, Minneapolis, USA, June 2007. IEEE Computer Society, Omnipress. Authorship 50-50.

[P8] J. Kostlivá, J. Čech, and R. Šára. Feasibility boundary in dense and semi-dense stereo matching. In *BenCOS 2007: CVPR Workshop Towards Benchmarking Automated Calibration, Orientation and Surface Reconstruction from Images*, Minneapolis, USA, June 2007. IEEE Computer Society, Omnipress. Authorship 34-33-33, Best Paper Award.

[P9] J. Čech and R. Šára. Complex correlation statistic for dense stereoscopic matching. In H. Kalviainen, J. Parkkinnen, and A. Kaarna, editors, *SCIA 2005: Proceedings of the 14th Scandinavian Conference on Image Analysis*, volume 3540 of *LNCS*, pages 598–608, Joensuu, Finland, June 2005. Springer-Verlag. Authorship 50-50.

[P10] J. Kostková, J. Čech, and R. Šára. Dense stereomatching algorithm performance for view prediction and structure reconstruction. In J. Bigun and T. Gustavsson, editors, *SCIA 2003: Proceedings of the 13th Scandinavian Conference on Image Analysis*, volume 2749 of *LNCS*, pages 101–107, Göteborg, Sweden, June 2003. Springer-Verlag. Authorship 33-34-33.

[P11] H. Zhang, J. Čech, R. Šára, F. Wu, and Z. Hu. A linear trinocular rectification method for accurate stereoscopic matching. In R. Harvey and J. A. Bangham, editors, *BMVC 2003: Proceedings of the 14th British Machine Vision Conference*, volume 1, pages 281–290, Norwich, UK, September 2003. British Machine Vision Association. Authorship 35-35-10-10-10.

[P12] V. Smutný, J. Čech, R. Šára, and T. Dostálová. Estimation of the temporomandibular joint trajectory by photogrammetry. In J. Jan, J. Kozumplík, and I. Provazník, editors, *Analysis of Biomedical Signals and Images, Proceedings of 16th Biennial International EURASIP Conference BIOSIGNAL 2002*, pages 271–273, Brno University of Technology, Brno, Czech Republic, June 2002. VUTIUM Press. Authorship 40-45-10-5.

[P13] V. Smutný, J. Čech, R. Šára, and T. Dostálová. Estimation of the temporomandibular joint position. In H. Wildenauer and W. Kropatsch, editors, *Proceedings of the CVWW'02*, pages 306–314, Bad Aussee, Austria, February 2002. Pattern Recognition & Image Processing Group, Vienna University of Technology. Authorship 40-40-10-10.

Technical reports:

[P14] J. Kostlivá, J. Čech, and R. Šára. ROC based evaluation of stereo algorithms. Research Report CTU–CMP–2007–08, Center for Machine Perception, K13133 FEE Czech Technical University, Prague, Czech Republic, March 2007. Authorship 34-33-33.

[P15] J. Čech. Towards accurate stereoscopic matching. Research Report CTU–CMP–2004–05, Center for Machine Perception, K13133 FEE Czech Technical University, Prague, Czech Republic, February 2004.

[P16] J. Čech and R. Šára. Efficient algorithms for computing correlation tables in stereoscopic vision. Research Report CTU–CMP–2004–11, Center for Machine Perception, K13133 FEE Czech Technical University, Prague, Czech Republic, October 2004. Authorship 50-50.

[P17] J. Kostková, J. Čech, and R. Šára. The CMP evaluation of stereo algorithms. Research Report CTU–CMP–2003–01, Center for Machine Perception, K333 FEE, Czech Technical University, Prague, Czech Republic, January 2003. Authorship 33-34-33.

[P18] J. Čech. Měření tvaru kloubní dráhy spodní čelisti. Master's thesis, Center for Machine Perception, K333 FEE, Czech Technical University, Prague, Czech Republic, February 2002. In Czech.

[P19] J. Čech. Nalezení optimálního pohledu kamery při měření kloubní dráhy dolní čelisti. Research Report CTU–CMP–2000–25, Center for Machine Perception, K333 FEE Czech Technical University, Prague, Czech Republic, December 2000. In Czech.

C Citations of author's work

[C1] L. Wang, H. Jin, and R. Yang. Search space reduction for MRF stereo. In *Proc. ECCV*, LNCS 5302, pages 576-588, 2008. Cites [P7].

[C2] S. Ivekovic and E. Trucco. Articulated 3-D modelling in a wide-baseline disparity space. In *Proc. 4th European Conference on Visual Media Production*, 2007. Cites [P7].

[C3] D. Martince and T. Pajdla. Robust rotation and translation estimation in multiview reconstruction. In *Proc. CVPR*, 2007. Cites [P7].

[C4] Z. Wei, X. Weisheng, and Y. Youling. Area harmony dominating rectification method for SIFT image matching. In *Proc. 8th International Conference on Electronic Measurement and Instruments (ICEMI'07)*, volume 2, pages 935–939, 2007. Cites [P3].

[C5] X.-Z. Li and G.-J. Zhang. Epipolar rectification in trinocular vision images. *Opto-Electronic Engineering*, 34(10):50-54, 2007. Cites [P3].

[C6] M. Shimizu and M. Okutomi. Multi-parameter simultaneous estimation on area-based matching. *International Journal of Computer Vision*, 67(3):327–342, 2006. Cites [P9].

[C7] C. Teutsch, D. Berndt, A. Sobotta, and S. Sperling. A flexible photogrammetric stereo vision system for capturing the 3D shape of extruded profiles. In P. S. Huang, editor, *Proc. Two- and Three-Dimensional Methods for Inspection and Metrology IV (Optics East 2006, October 1-4, 2006, Boston, MA, USA)*, volume 6382, pages 63820M.1–63820M.9. SPIE, 2006. Cites [P9].

[C8] M. Heinrichs and V. Rodehorst. Trinocular rectification for various camera setups. In *Proc. Symposium of ISPRS Commission III - Photogrammetric Computer Vision (PCV'06)*, pages 43–48, 2006. Cites [P11].

[C9] F. Kangni and R. Laganière. Projective retification of image triplets from the fundamental matrix. In *Proc. IEEE International Conference on Acoustic, Speech and Signal Processing*, volume 2, pages 505–508, 2006. Cites [P11].

[C10] P. Rongjiang and M. Xiangxu. Robust estimation of rotation axis of pottery sherds. *Journal of Computer-Aided Design and Computer Graphics*, 17(15):2508–2511, 2005. Cites [P3].

[C11] B. Bocquillon, S. Chambon, and A. Crouzil. Segmentation semi-automatique en plans pour la génération de cartes denses de disparités. Technical Report IRIT/2005-23-R, IRIT, Université Paul Sabatier, Toulouse, France, 2005. Cites [P10].

Citations [C1, C3, C6, C7, C9] are recorded in SCI database.

D Other community acceptance

We are currently registering over 300 downloads of our Matlab Toolbox which implements the stereoscopic matching based on Growing Correspondence Seeds (GCS), originally published in [P7]. The toolbox is available for non-commercial use at `http://cmp.felk.cvut.cz/~cechj/GCS/`.

A non-exclusive licence of the GCS software was bought by a Canadian company Feeling Software Inc.