Estimation and Image Segmentation of a Sparse Disparity Map for 3D Reconstruction

Patrik Kamencay, Martin Breznan, Roman Jarina, Martina Zachariasova
Dept. of Telecommunications and Multimedia
University of Zilina
Zilina, Slovakia
patrik.kamencay@fel.uniza.sk

***Abstract*—In this paper, a stereo matching algorithm based on image segments is presented. We propose the hybrid segmentation algorithm that is based on a combination of the Belief Propagation and K-Means algorithms with aim to refine the final sparse disparity map by using a stereo pair of images. Firstly, a color based segmentation method is applied for segmenting the left image of the input stereo pair (reference image) into regions. The aim of the segmentation is to simplify representation of the image into the form that is easier to analyze and is able to locate objects in images. Secondly, results of the segmentation are used as an input of the SIFT-SAD matching method to determine the disparity estimate of each image pixel. This matching algorithm is proposed by combining Scale Invariant Feature Transform (SIFT) with the Sum of Absolute Difference (SAD). Finally, the comparisons between the three robust feature detection methods SIFT, Affine SIFT (ASIFT) and Speeded Up Robust Features (SURF) are presented. The obtained experimental results demonstrate that the final depth map can be obtained by application of segment disparities to the original images.**

***Index Terms*—Belief Propagation, K-Means, SIFT, ASIFT, SURF, disparity map.**

I. Introduction

This paper describes a set of algorithms for structure, motion automatic recovery and visualization of a 3D image from a sequence of 2D images. The important step to perform this goal is matching of corresponding pixels in the different views to estimate the depth map. The depth of an image pixel is the distance of the corresponding world point from the camera center. Detecting objects, estimating their pose, geometric properties and recovering 3D shape information are a critical problem in many vision and stereo computer vision application domains such as robotics applications, high level visual scene understanding, activity recognition, and object modelling [1]. The structure and motion recovery system follows a natural progression, comprising the following phases:

- image acquisition,
- feature matching using SIFT descriptor,
- image segmentation,
- feature detection using SIFT-SAD algorithm,
- stereo geometry and image rectification,
- estimation of disparity and final depth map.

A classical problem of stereo computer vision is the extraction of 3D information from stereo views of a scene. To solve this problem, knowledge of view properties and feature point between views is needed. However, finding these points is notoriously hard to do for natural scenes. The fundamental idea behind stereo computer vision is the difference in position of a unique 3D point in two different images. As the object moves closer to the cameras, the relative position of object will change, and the positions in each image will move away from each other. In this way, is possible to calculate the distance of an object, by calculating its relative positioning in the two images. This distance between the same objects in two images is known as disparity [1]. Disparity map computation is one of the key problems in 3D computer vision.

This paper employed a new feature projection approach based on SIFT-SAD method using hybrid segmentation algorithm. A comparison between these two different approaches for the image segmentation (K-Means and Belief Propagation) is described in [2], [3].

The outline of the paper is as follows. In the next section, an overview of image segmentation methods (K-Means and Belief propagation) is introduced. Disparity map calculation from corresponding points is described in Section III. Finally the experiment results and architecture of reconstruction algorithm are introduced in Section V and brief summary is discussed in Section VI.

II. Image Segmentation

The main goal of the image segmentation is split the entire image into set of segments that cover image. In this chapter, K-Means and Belief Propagation segmentation algorithms will be presented.

A. K-Means segmentation algorithm

K-means algorithm is statistical clustering algorithm. Data clustering is method that creates groups of objects (clusters). K-means algorithm is based upon the index of similarity or dissimilarity between pairs of data components [3]. This type of algorithm is popular for simplicity, implementation and it is commonly used for grouping pixels in images. But this algorithm has three basic disadvantages [3], [5]:

- *K*, the number of clusters must be determined.

- Different initial conditions produce different results.
- The data far away from center pull the centers away from optimum location.

Let $X = \{x_1, x_2, \ldots, x_n\} \subset R^p$ be a finite set of data where N is the number of data items and R^p is p-dimensional Euclidean space [5]. Let V_{KN} be the set of matrices ($K \times N$, $2 \leq K < N$), where K is number of clusters. A K partition of X is defined:

$$M_K = \left\{ U \in V_{K,N} \left| \begin{array}{l} u_{ik} \in \{0,1\}, \forall i,k; \\ \sum_{i=1}^{K} u_{ik} = 1, \forall k \end{array} \right. \right\}, \quad (1)$$

where $u_{ik} = 1$ denotes that component x_k belongs to cluster i, $u_{jk} = 0$ denotes that component x_k is out of the cluster j. The objective function J_K is:

$$J_K(U,K) = \sum_{i=1}^{K} \sum_{k=1}^{N} u_{ik} d_{ik}^2, \quad (2)$$

where d_{ik} is Euclidean distance between component x_k and cluster v_i:

$$d_{ik} = \|x_k - v_i\| \quad (3)$$

Optimal K clusters of X is produced by minimization of objective function $J_K(U, V)$ [3], [5].

B. Belief Propagation

Belief propagation is an iterative inference algorithm for graphical models such as MRF which is based on a message passing principle that propagates messages in the network [2], [4].

The above model contains only pairwise cliques, and the joint probability over the 3D volume is

$$P = \prod_{i \neq j} \psi_{i,j}(s_i, s_j) \prod_k \varphi_k(s_k, d_k), \quad (4)$$

where s_j and d_j represent state node and data node separately. Ψ is the state transition function between a pair of different hidden state nodes and ϕ is the measurement function between the hidden state node and observed data node. N represents the total number of state or data nodes in the 3D volume. Under the squared loss function, the best estimate for node s_j is the mean of the posterior marginal probability (minimum mean squared error estimate, MSE estimate):

$$s_{jMSE} = \sum_{s_j} s_j \sum_{s_i, i \neq j} P(s_1, .., s_N, d_1, .., d_N), \quad (5)$$

where the inner sum gives the marginal distribution of s_j [1], [2], [4].

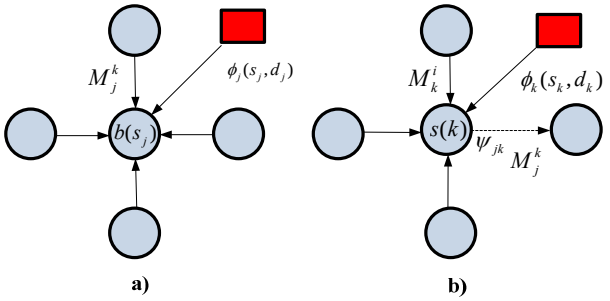

Fig. 1. (a) Computing belief, (b) Computing message.

Since the joint probability involves all the hidden state nodes and data nodes in the 3D volume, it is hard to compute the MSE estimate based on the implicit multivariable probability distribution. However belief propagation messages are effective to compute the MSE estimate recursively. Each hidden state node has a belief, which is a probability distribution defining the node's motion likelihood. Thus the MSE estimate of one node is computed as:

$$s_{jMSE} = \sum_{s_j} s_j b(s_j), \quad (6)$$

where

$$b(s_j) = \varphi_j(s_j, d_j) \prod_{k \in Neighbor(j)} M_j^k, \quad (7)$$

is the belief at node s_j and k runs over all neighboring hidden state nodes of node s_j. The belief at node s_j is the product of all the incoming messages M and the local observed data message ($\phi_j(s_j, d_j)$). The computation is shown in Fig. 1 (a). The passed messages specify what distribution each node thinks its neighbors should have. Fig. 1 (b) shows how to compute the message from node s_k to s_j :

$$M_j^k = \sum_{s_k} \psi_{jk}(s_j, s_k) b(s_k). \quad (8)$$

After multiplying all the incoming messages M from neighbouring nodes (except from the node s_j) and the observed data message ($\phi_k(s_k, d_k)$), the product is evolved from the message-sender to the message-receiver by transition function $\psi_{jk}(s_j, s_k)$ [2], [4].

C. Hybrid segmentation

Hybrid methods are created by combining two or more image segmentation algorithms. In our image analysis, a hybrid algorithm, which is produced by the combination Belief Propagation [2], [4] and K-Means [3], [5] is used. This approach brings together the advantages of both segmentation algorithms. K-Means is quick and Belief Propagation is very accurate segmentation. Diagram of hybrid segmentation algorithm is shown in Fig. 2.

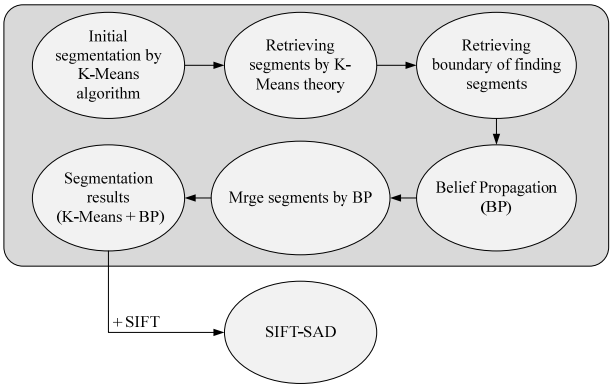

Fig. 2. Hybrid segmentation algorithm.

In first step, the hybrid algorithm is applied to image filtering using Mean Shift following the image dividing into segments by applying the K-Means. In the second step, the image is split into segments using K-Means algorithm. In final step, the similar small segments are combined into bigger segments, through Belief Propagation [10].

III. Feature Calculation

The Feature matching algorithm improves precision of disparity calculation. This kind of algorithms extracts object's suitable features in the 3D scene, for example, segments of edges or contours in the left and right stereo images [8]. In the following stage, disparity map is calculated from corresponding points of the features.

A reconstruction of the disparity map from the left and right stereo pair is known as the stereo matching algorithm. The detection feature points must be matched. There exist several matching techniques based on various algorithms, e.g. Correlation (C), Normalized Cross Correlation (NCC), Sum of Squared Differences (SSD) and Sum of Absolute Differences (SAD) algorithms. In our case we used SAD matching algorithm [6].

A. SIFT descriptor

Scale Invariant Feature Transform (SIFT) is a local descriptor of image features insensitive to illuminant and other variants that is usually used as sparse feature representation [7]. SIFT features are features extracted from images to help in reliable matching between different views of the same object [6]. The extracted features are invariant to scale and orientation, and are highly distinctive of the image. They are extracted in four steps. The first step computes the locations of potential interest points in the image by detecting the maxima and minima of a set of Difference of Gaussian (DoG) filters applied at different scales all over the image. Then, these locations are refined by discarding points of low contrast. An orientation is then assigned to each key point based on local image features. Finally, a local feature descriptor is computed at each key point. This descriptor is based on the local image gradient, transformed according to the orientation of the key point to provide orientation invariance. Every feature is a vector of dimension 128 distinctively identifying the neighbourhood around the key point [7].

B. Sum of Absolute Differences (SAD)

The SAD is widely used metric for block matching in stereo images. It works by taking the absolute value of the difference between each pixel in the original block and the corresponding pixel in the block being used for comparison. These differences are summed over the block to create a simple metric of block similarity, the $L1$ norm of the difference image [8]. The cost function $C(p)$ on the basis of SAD is computed as follows:

$$C(p) = \sum_{(p')\in W(p)}^{N} \left| I_L(p') - I_R(p') \right|, \tag{9}$$

where p is the reference pixel at which the SAD is computed and p' represent the pixels belonging to the neighbourhood of the pixel p. $W(p)$ is square window that surrounds the position (p_x, p_y) of the pixel. The minimum difference value over the frame indicates the best matching pixel, and position of the minimum defines the disparity of the actual pixel [6], [8].

C. SIFT-SAD algorithm

The final performance of stereo matching algorithms depends on the choice of matching cost. In our experiment we proposed SIFT-SAD matching method as matching cost. SIFT descriptor delivers most of local gradient information and SAD provides local intensity information. SIFT-SAD consists of two parts. Firstly, we get the $L1$ distance of SIFT between pixel p in the left image and $p+d_p$ in the right image [6], [8].

$$D_{SIFT}(d_p) = \left\| x_L(p) - x_R(p + d_p) \right\|, \tag{10}$$

where d_p is the disparity of pixel, $\|x_L(p)\text{-}x_R(p+d_p)\|$ is the $L1$ distance. Next, we define SAD matching cost as:

$$D_{SAD}(d_p) = \exp\left(-SAD\left(p, p + d_p\right)\right), \tag{11}$$

where $SAD(p, p+d_p)$ is the SAD score in a square neighbourhood searching window. Our algorithm computes the disparity for all pixels with window size dimension at square of 9x9 pixels. Finally, we use one dimensional Gaussian weight with a scale factor s to get the matching cost. The underlying assumption is that if a minimum corresponds to the true surface, the neighbouring pixels should have near values at a similar depth [8].

IV. Evaluation Criterion

The criterion used for comparing image segmentation algorithms presented in this article, is based on computing precision, recall and *F1*. These three parameters determine the algorithms efficiency by comparing boundaries their segments. Each of the algorithms is compared with segmentation by a human. Based on this comparison, precision, recall and *F1* are computed. The definition of precision, recall and *F1* is given by:

$$P = \frac{C}{C+F} \cdot 100\%, \tag{12}$$

$$R = \frac{C}{C+M} \cdot 100\%, \tag{13}$$

where C is the number of correct detected pixels that belongs to boundary, F is the number of false detected pixels and M is the number of not detected pixels.

F1 is combined measure from precision and recall. It is in high values if both precision and recall have high values and on the other hand, if one of them has low value, the value of the F1 is going down. The definition of F1 is given by:

$$F1 = \frac{2PR}{P+R} \cdot 100\%. \tag{14}$$

V. Experimental Results

In this section, some of the obtained experimental results rectifying, matching points and generating sparse disparity map will be presented. The proposed architecture (see Fig. 3) has been tested on two input real images. This proposed algorithm based on the combination of SAD stereo matching algorithm with SIFT descriptor is faster, since a small portion of whole left and right images pixels are used for matching. In this experiment, we segment the reference image (in our case, the left image) using hybrid segmentation method. Then, for each segment we look at the associated pixel disparities.

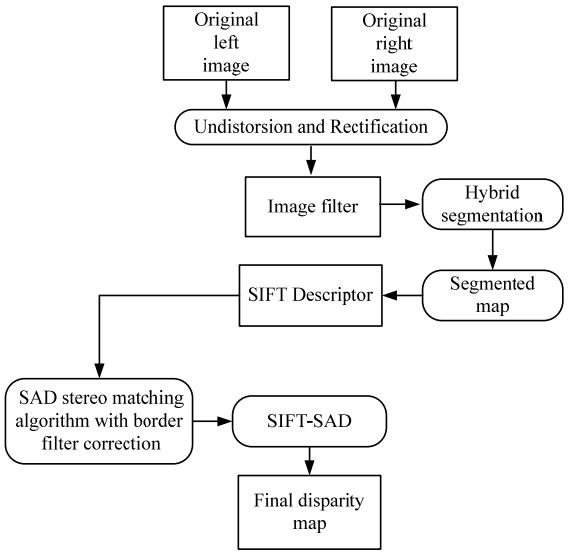

Fig. 3. Architecture of reconstruction algorithm.

First, the edges are extracted using Harris method due to its good performance. Harris corner detector is a suitable starting point for the computation of positions of scale [12]. Next step is image rectification. It is transformation which makes pairs of conjugate epipolar lines become collinear and parallel to the horizontal axis (baseline). For the epipolar rectified images pair, each point in the left image lies on the same horizontal scan line as in the right image. This approach is used to reduce a search space for disparity map estimation algorithm. Then, we apply image filtering by Mean Shift filter. This step is very useful for noise removing, smoothing and image segmentation [13], [14]. For each pixel of the image, the set of neighboring pixels is determined. Let X_i be the input and Y_i filtered image, where $i = 1, 2, \ldots, n$. The filtering algorithm comprises of the following steps:

- Initialize $j = 1$ and $y_{i,1} = p_i$.
- Compute through the Mean Shift the mode where the pixel converges.
- Store the component of the gray level of the calculated value $Z_i = (x_i, y_{i,c})$ at Z_i , where x_i is the spatial component and $y_{i,c}$ is the range component.

After filtration, the filtered image is split into segments using K-Means segmentation algorithm. In the next step, the small segments are merged together to the most similar adjacent segments. Next, matching is performed, where a sparse disparity map is obtained. The match points can be obtained using SAD approach along the epipolar line [15]. The disparity map codifies the distance between the object and the camera - closer points will have maximal disparity and farther points will get zero disparity [16], [17]. Finally, we have integrated the segmentation algorithm K-Means with the SIFT-SAD method. This method consists of two parts: SIFT part and SAD part. As already stated, our experimental results proved that the SIFT descriptor is a very robust and reliable representation for the local neighborhood of an image point. This proposed approach is able to produce highly accurate disparity map.

Quality of disparity map is represented as percentage of pixels with disparity errors (bad matching pixels) [17]:

$$P = \frac{1}{X * Y} \sum_{i=1}^{X} \sum_{j=1}^{Y} \left(\left| d_C(i,j) - d_T(i,j) \right| \right), \tag{15}$$

where $X*Y$ represent the size of the image, d_C is the computed disparity map of the test image and d_T is the truth disparity map [10].

$$d_T = \frac{fBI_{RES}}{D_T h}, \tag{16}$$

where D_T is ground truth depth map, h is height from the ground plane, D_T*h is ground truth distance, B is baseline between the cameras, I_{RES} is image resolution and f is focal length.

Table I shows summary of overall performance. We compared performances obtained by the proposed method SIFT-SAD with those obtained by three common algorithms (SIFT, ASIFT and SURF - Speeded Up Robust features) [2], [6], [9]. Approximately 92 percent of the disparity values were

found correctly for our proposed algorithm. The final disparity map is labeled as correct if it is within one pixel of the correct disparity. The ground truth disparity map [17] is the inverse of the ground truth distance scale by the image resolution and the focal length [15], [18]. Equation (16) shows how to calculate the ground truth disparity map from the depth map. The depth map is a 16 bit map with values ranging from 0 to 1 where the ground plane was at *D=1* and the cameras were at *D=0*. *D* is distance of object from the camera.

TABLE I. THE PERCENTAGE OF DISPARITIES FOUND CORRECTLY, DISPARITY ERROR AND THE DETECTED OCCLUSION THAT ARE CORRECT [18]

	SIFT	***ASIFT***	***SURF***	***SIFT-SAD***
Disparity Correct [%]	86,69	82,07	89,78	92,35
Disparity Error [%]	13,31	17,93	10,22	7,65
Occlusion Correct	67,45	65,32	72,76	72,03

In this experiment three segmentation algorithms were compared using automatic algorithm evaluating the precision of segmentation, as is shown in Tab. II. This plays important role for two reasons: (1) it can be placed into a feedback loop to enforce another run of segmentation algorithm that may include more sophisticated steps for high precision segmentation and (2) the outcome of this evaluation can be treated as a quality factor and thus can be used to design a quality driven adaptive recognition system.

TABLE II. BEST RESULTS OF IMAGE SEGMENTATION ALGORITHM

Image Segmentation	***Precision [%]***	***Recall [%]***	***F1 [%]***
Belief Propagation	55,34	19,47	21,03
K-Means	43,27	15,13	17,56
Hybrid Segmentation	61,49	25,09	27,52

TABLE III. PARAMETERS USED IN HYBRID ALGORITHM

Parameter	***Set value***
s	5
S	50
Min_sh	1
Max_sh	40

In the next experiment, the hybrid segmentation algorithm implementation to proposed algorithm was tested. The set up parameters of used hybrid segmentation algorithm are shown in Tab. III. Spatial resolution parameter *s* affects smoothing and connectivity of segments. Moreover, parameter *S* is a size of the smallest segment, *Min_sh* is minimum and *Max_sh* is maximum shift of the pixels. Finally the hybrid segmentation algorithm was integrated with SIFT-SAD method. The results show, that SIFT-SAD method with use of image segmentation achieved better results. All the experiments were practiced on pictures from 101 object images database [11]. The presented algorithms were implemented in MATLAB.

The experimental results are presented in Table IV, where *S* is segmented and *N* is non-segmented image. Moreover, in both tables a conformity as a percentage of images before and after segmentation as well as the total number of matches found are shown. The accuracy of matching was computed by simply formula:

$$P_T = \frac{K_M}{K_A}.100, \tag{17}$$

where K_A is the number of all matches found, K_M is the number of truly matches found in the images [9].

TABLE IV. RESULTS OF SIFT-SAD METHOD USED WITH AND WITHOUT HYBRID SEGMENTATION ALGORITHM

Class title	***Number of key points***	K_A	K_M	P_T	
Beaver	1486	35	27	77,14	**S**
	2713	49	24	48,98	**N**
Tick	1887	327	315	96,33	**S**
	3399	291	232	79,73	**N**
Flamingo	1116	41	33	80,49	**S**
	1732	75	41	54,66	**N**
Elephant	1743	34	21	61,76	**S**
	3645	47	19	40,43	**N**
Dog	1035	132	113	85,61	**S**
	1364	214	168	78,51	**N**

The results shows, that the number of keypoints for image description for segmented images (S) was decreased and the accuracy of keypoints matching was increased.

Finally, the four methods (SIFT, ASIFT, SURF, SIFT-SAD) are compared. This all methods are based on combination OpenCV and MATLAB. We use the same dataset, which includes the general deformations, such as view, illumination and rotation changes. Time evaluation is a relative result, which only shows the tendency of the four methods' time cost. There are factors that influenced on the results such as the size and quality of the image, image types (e.g. scenery or texture), and the parameters of the algorithm (e.g. the distance ratio) [6].

TABLE V. PROCESSING TIME COMPARISON

	SIFT	***ASIFT***	***SURF***	***SIFT-SAD***
Total matches	125	135	89	312
Total time [s]	6,82	5,07	2,78	4,95
15 matches time [s]	4,15	4,72	2,07	4,31

In this part of the experiment we uses 101 object image dataset, whose sizes are all 300 x 240 pixels. The parameters of the four algorithms are the same settings according to the original paper. Time is counted for the complete processing which includes feature detecting and matching. Total matching time (see Tab. V) is the computational time of finding all matches. The computational time for SIFT descriptor was approximately 7 seconds and for proposed SIFT-SAD method 5 seconds, respectively. Tab. V shown that SURF is fastest one, SIFT-SAD is slower but it detects so many key points and finds most matches. Furthermore, the proposed method is noise insensitive.

The comparison among the three algorithms, the experimental results show that the proposed SIFT-SAD method is far efficient than SIFT, ASIFT or SURF algorithm.

VI. Conclusion

The method for reconstructing a 3D scene from two input images was presented. We mentioned some manners allowing a three - dimensional reconstruction of picture or object in this article. The proposed system is based on 3D reconstruction solution using stereo images. This system works with common cameras. The applications of these methods of 3D picture processing are very useful in sphere of medicine, for example detection and identification of tumor in brain and also in other branches as physics, astronomy, biology or geography. Future task we could speed up computation time and improve precision of Belief propagation algorithm.

Acknowledgment

This paper has been supported by the Slovak Science project Grant Agency, Project No. 26220220407 "Competence center for research and development of diagnostics and therapy of cancer".

References

[1] J. S. Yedida, W. T. Freeman, Y. Weiss, "Understanding belief propagation and it is generalizations", *Exploring Artificial Intelligence in the New Millennium*, Chap. 8, pp. 239-236, January 2003.

[2] J. Sun, N. N. Zhang, H. Y. Shum, "Stereo matching using belief propagation". *In IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(7):787-800, 2003.

[3] L. Dong, P. Ogunbona, W. Li, G. Yu, L. Fan, G. Zheng, "A fast algorithm for color image segmentation". *First International Conference on Innovative Computing*, Information and Control, 2006.

[4] Z. Yin, R. Collins, "Belief Propagation in a 3D Spatio - temporal MRF for moving object detection", Department of Computer Science and Engineering, The Pennsylvania State University, University Park, PA, 2007.

[5] M. Benco, R. Hudec, "The advanced image segmentation techniques for broadly useful retrieval in large image database", *NSSS*, Tatranské Zruby, Slovakia, pp. 40-44, May 2006, ISBN 978-80-8040-344-7.

[6] L. Juan, O. Gwun, "A Comparison of SIFT, PCA-SIFT and SURF," *International Journal of Image Processing (IJIP)*, Vol. 3, No. 4., 2009, pp. 143-152.

[7] D. G. Lowe, "Distinctive image feature from scale-invariant keypoints," *International Journal of Computer Vision (IJCV)*, Vol. 2, No. 60, 2004, pp. 91-110.

[8] A. Kuhl, "A Comparison of Stereo Matching Algorithms for Mobile Robots," *Centre for Intelligent Information Processing System.* 2005, The University of Western Australia, pp. 4-24.

[9] M. Zachariasova, R. Hudec, M. Benco, P. Kamencay, "The object recognition based on Scale-Invariant Feature Transform and Hybrid Segmentation," *Proceedings of the 9th International Conference ELEKTRO 2012.* May 21-22, CD ISBN 978-1-4673-1178-6.

[10] P. Kamencay, M. Breznan, R. Jarina, P. Lukac, M. Zachariasova, "Improved depth map estimation from stereo images based on hybrod method," *The Radioengineering journal.* Vol. 21, No. 1, April 2012, ISSN 1210-2512.

[11] Computational Vision at Caltech 101 [Online] Cited 2012-07-17. Available at: http:://vision.caltech.edu/Image_Datasets/.

[12] K. Mikolajczyk, C. Schmid, "Scale and Affine Invariant Interest Point Detectors". *International Journal of Computer Vision (IJCV)*. vol. 60, no. 1, 2008, p. 63 - 86.

[13] R. Hudec, M. Benco, J. Krajcovic, J., "SD LMS L-filters for filtration of gray level images in time-spatial domain based on GLCM features." *Advances in Electrical and Electronic Engineering*, Zilina (Slovakia), 2008, vol. 7, No. 1-2, p. 51 - 54.

[14] R. Hudec, "*Adaptive order-statistics L-filters*." 1st ed. University of Zilina: EDIS Press, 155 pages, ISBN 978-80-554-0248-2, 2011 (in Slovak).

[15] M. Shimizu, M. Okutomi, "Calibration and rectification for reflection stereo." *Computer Vision and Pattern Recognition CVPR. IEEE Conference*. 2008, p. 1 - 8.

[16] D. Scharstein, R. Szeliski, "A taxonomy and evaluation of dense two frame stereo correspondence algorithms." *International Journal of Computer Vision,* 2002, p. 7 - 42.

[17] R. Märtin, "Towards Ground Truth Disparity Maps of Natural Scenes: Combining Laser Range Scans and Stereo Images". *Publications of the Institute of Cognitive Science*, 2006, vol. 23-2010.

[18] R. Queiroz, M. Cohen, J. L. Moreira, A. Braun, J. C. Jacques, S. R. Musse, "Generation Facial Ground Truth with Synthetic Faces". *Graphics, Patterns and Images (SIBGRAPI)*. 2010, p. 25 - 31.