Depth estimation based on Maximization of A posteriori Probability

Olgierd Stankiewicz, Krzysztof Wegner, and Marek Domanski

Chair of Multimedia Telecommunications and Microelectronics,
Poznan University of Technology, Poland
ostank@multimedia.edu.pl

Abstract. This paper presents a proposal of depth estimation method which employs empirical modeling of cost function based on Maximization of A posteriori Probability (MAP) rule. The proposed method allows for unsupervised depth estimation without a need for usage of arbitrary settings or control parameters, like Smoothing Coefficient in Depth Estimation Reference Software (DERS), which was used as a reference. The attained quality of generated depth maps is comparable to a case when supervised depth estimation is used, and such parameters are manually optimized. In the case when sub-optimal settings of control parameters in supervised depth estimation with DERS is used, the proposed method provides gains of about 2.8dB measured in average PSNR quality of virtual views synthesized with the use of estimated depth maps in the tested sequence set.

1 Introduction

Depth map is a practical format of 3D representation of a scene [1]. A common method to obtain depth maps is to estimate them algorithmically from a video. Although the first works on depth estimation go back to 1950's, the current state of the art is still far away from satisfying level in many applications, especially in case of new generation of 3D video systems [2].

The basic principle of algorithmic depth estimation is finding correspondence between features in two (or more) views of the same 3D scene [1]. Estimation of depth map requires finding correspondence (disparity) between all pixels in one view and pixels in other views. This problem is computationally expensive and is typically solved by employing generic optimization algorithms [4] like Belief Propagation or Graph Cuts. In order to do so, a cost function over a depth map is formulated. Such function is often related to as "energy", "goal function" or "performance index" in other optimization applications.

In this paper we present a novel, theoretically founded approach to depth estimation which employs Maximum A posteriori Probability (MAP) rule for modeling of the cost function used in optimization algorithms. The proposal is presented along with a method for estimation of parameters of such model.

2 State of the art

In depth estimation, typically the cost function (denoted $Fitness$) is modeled as a sum of two sub-functions: $DataCost$ and $TransitionCost$ for each pixel:

$$Fitness = \sum_{p} DataCost_p\left(d_q\right) + \sum_{q \in (neighborhood\ of\ p)} TransitionCost_{p \to q}\left(d_p, d_q\right) \quad (1)$$

where p is a particular pixel in the considered depth map, q is some pixel (point) in the neighborhood of pixel p in the same view, d_p and d_q are assumed depth values for pixels p and q respectively. The terms $DataCost_p\left(d_q\right)$ and $TransitionCost_{p \to q}\left(d_p, d_q\right)$ are functions described below.

$DataCost_p\left(d_q\right)$ models the direct correspondence between pixels and expresses how given pixel p is similar to pixel in one of other images pointed by its depth d_p. The higher is the difference between those pixels, the higher is the value of $DataCost_p\left(d_q\right)$. The most commonly $DataCost$ is defined in terms of energy related to similarity metrics between fragments of images, calculated in pixels or blocks. Typically, Sum of Absolute Differences (SAD) [5] or Sum of Squared Differences (SSD) [6][7] metrics are used. Some state-of-the-art works which relate to $DataCost$ function propose usage of "rank" or "census" [8] for calculation of better similarity metric. Work [9] proposes a more advanced approach, where mixture of various similarity metrics is incorporated in order to attain better quality in depth estimation but theoretical foundations are missing. In paper [7] authors provide a derivation of $FitCost$ function based on MAP assumptions. Unfortunately, the work omits the consequences of this derivation related to $DataCost$ which limited the work to consideration of gaussian model (corresponding to Sum of Squared differences energy formulation). No verification is provided, whether such assumptions are correct. Similarly, work [10] employs posteriori probability for modeling of $Fitness$. Authors consider a more advanced model for $DataCost$ which incorporates Generalized gaussian-like model with arbitrary power exponent. Also this work does not provide any verification of whether taken assumptions are correct, apart from theoretical considerations. In work [11] authors have proposed usage of truncated-linear $DataCost$ function which actually corresponds to Absolute Difference similarity metric, additionally saturated, so that it does not exceed some given maximal level. Apart from the concept being very scientifically interesting and giving promising results, the authors have not supported their proposal with empirical data verifying their assumptions. In work [12] authors thoughtfully analyze probabilistic model of correspondence in 3D space. Instead of MAP rule, a different approach for evaluating entropy and mutual information, called EMMA, is proposed. Unfortunately, the method is presented in context of 3D modeling and not depth map estimation itself which disallows comparison with other state-of-the-art methods.

The second component of $Fitness$ function, $TransitionCost_{p \to q}\left(d_p, d_q\right)$, penalizes depth maps that are not smooth. Its role is regularization of the resultant depth map. The higher are the differences between depth d_p of pixel p and depth values d_q of all neighboring pixels q, the higher is the value of

$TransitionCost_{p\to q}(d_p, d_q)$. Typically, $TransitionCost_{p\to q}(d_p, d_q)$ is defined independently from particular pixel positions p and q and thus can be simplified to $TransitionCost(d_p, d_q)$. Also, very often, $TransitionCost$ is not defined as a function of d_p and d_q independently, but as a function of their absolute difference $|d_p - d_q|$ only: $TransitionCost(|d_p - d_q|)$. Among the most commonly known models for *TransitionCost* function of this type there are Potts model [13], Linear model [14][15] and Truncated-linear model [11].

In general, $TransitionCost$ functions incorporate some sort of constant parameters, like the penalty value in Potts Model or the slope of penalty segment in Linear Model. The main purpose of such constant parameters is to provide weighting of $TransitionCost$ related to *DataCost* function (to which it is added to formulate $Fitness$ function (1)). The most commonly, such a weight is called "Smoothing Coefficient" as its value sets how much depth maps that are not smooth are penalized by $Fitness$ function. The use of small values of Smoothing Coefficient results in sharp but noisy depth maps. The use of large values results in very smooth, even blurred depth maps. The selection of particulac value is typically done manually (the depth estimation is thus supervised) which is an important problem in practical use of depth estimation methods in applications, where unsupervised operation is expected.

All of the mentioned models (Potts, linear and truncated-linear) are widely used because they are simple and can be efficiently implemented in optimization algorithms like BP of GC algorithms. Unfortunately, the use of a specific model is rarely justified by scientific reasons. E.g. in work [11], authors have proposed usage of truncated-linear-shaped $TransitionCost$ function for depth estimation and have compared it against other state-of-the-art techniques. Although the results are promising, the foundations of the proposal are not given. In papers [7][10] authors consider derivation of $TransitionCost$ function based on MAP rule, similar to the approach in this paper. Markov Random Field model for stereoscopic depth estimation is formulated by means of BP algorithm. Unfortunately, the work proposes only an approximation of $TransitionCost$ function.

To summarize, there is lack of works which provide theoretical analysis of application of Maximum A posteriori Probability (MAP) optimization rule to formulate *DataCost* and $TransitionCost$ functions for depth estimation, along with empirical experimentation which would support formulation of such theoretical models. This lack is a motivation of this work.

3 Proposed cost function derivation based on MAP

Below we provide derivation of $Fitness$ function based on Maximum A posteriori Probability rule. First, a theoretical formulation for depth map estimation based on Maximum A posteriori Probability (MAP) optimization rule is presented. It is shown what are the assumptions required in order to attain classically used Absolute Differences [5] or Squared Differences [6][7] pixel similarity metrics in formulation of *DataCost* function. Then, similarly, a formulation of $TransitionCost$ function is proposed on the basis of a probabilistic model.

3.1 DataCost component

Let us consider depth estimation in a case of two identical cameras which are perfectly horizontally aligned with parallel optical axes. The views are rectified and the distortions are assumed to be removed [16]. Therefore, epipolar lines (along which correspondence search is performed) are aligned with horizontal rows in the images. Images from the left view $L_{x,y}$ and from the right view $R_{x,y}$ have the same widths W and the same heights H.

For given row y of pixels in both views, observed are pixel luminance values in the left view $(L_{1,y}, \cdots, L_{W,y})$ and in the right view $(R_{1,y}, \cdots, R_{W,y})$, both indexed from row 1 to row H. All of these are random variables are considered to have been observed and thus they constitute our a posteriori observation set.

We search for depth value $d_{x,y}$ for each pixel at coordinates x, y (in the right view) which would maximize probability $p\,(d_{x,y})$ under the condition of a posteriori observations of luminance values in both views. This probability will be demarked as $p_{x,y,d}$:

$$p_{x,y,d} \equiv p\,(d_{x,y} \mid (L_{1,y}, \cdots, L_{W,y}, R_{1,y}, \cdots, R_{W,y})) \tag{2}$$

where $\mid (L_{1,y}, \cdots, R_{W,y})$ is overall conditional expression of observation of luminance values. Therefore MAP rule for selecting optimal depth value $d^*_{x,y}$ is:

$$d^*_{x,y} = max_{arg\,d}\,(p_{x,y,d}) \tag{3}$$

In order to allow the depth estimation algorithm to use the MAP rule (3), the term $p_{x,y,d}$ has to be modeled basing solely on values that are known after the observation (a posteriori), e.g. luminance values in both of the views $L_{1\cdots W,y}$ and $R_{1\cdots W,y}$. Thus, with the use of the Bayes rule we will transform equation (2). Then, by rearrangement of $(L_{1,y}, \cdots, L_{W,y}, R_{1,y}, \cdots, R_{W,y}) \mid d_{x,y}$ term for each luminance variable separately (e.g. $L_{1,y}$), we get:

$$p_{x,y,d} = \frac{p\,(L_{1,y} \mid d_{x,y}, \cdots, L_{W,y} \mid d_{x,y}, R_{1,y} \mid d_{x,y}, \cdots, R_{W,y} \mid d_{x,y}) \cdot p\,(d_{x,y})}{p\,(L_{1,y}, \cdots, L_{W,y}, R_{1,y}, \cdots, R_{W,y})} \tag{4}$$

Assumed is presence of noise which has independent realizations in each of the views. Thus, pixel luminance values in the left view $L_{1,y}, \cdots, L_{W,y}$ are independent from each other and the same in the right view. This holds true for all terms in the denominator of (4), specifically also for the sought pair of pixels matched by depth $d_{x,y}$, as the denominator does not consider any specific matching or correspondence of pixels, as those probabilities are not conditional with respect to $d_{x,y}$. Therefore, we can simplify the denominator of (4) as $\prod_{l=1..W} p\,(L_{l,y}) \cdot \prod_{r=1..W} p\,(R_{r,y})$. A similar simplification could be done in the case of the nominator of (4), but here, on the contrary, probabilities are conditional because they are considered under the condition of occurrence of $d_{x,y}$. Such condition of $d_{x,y}$ means that in the given pixel (x, y), for which we calculate $p_{x,y,d}$, a depth value $d_{x,y}$ is assumed, so that two pixels, with coordinates l in the left and r in the right view, correspond to each other through depth $d_{x,y}$. For the

sake of brevity lets assume that $x + d_{x,y}$ operation will represent corresponding coordinate (just like $d_{x,y}$ would be direct disparity value), thus:

$$r = x,\ l = x + d_{x,y} \tag{5}$$

x expresses the coordinate in the right view for which $d_{x,y}$ is considered. Such pair of pixels is not independent, and therefore probabilities of their luminance values $p\left(L_{l,y}\right)$ and $p\left(R_{r,y}\right)$ cannot be simplified. For other pairs of pixels (not corresponding to each other) random variables describing their luminance values are independent. Therefore, we can express $p_{x,y,d}$ from (4) as:

$$p_{x,y,d} = \frac{\prod_{\substack{l=1..W,\\ l\neq x+d_{x,y}}} p\left(L_{l,y} \mid d_{x,y}\right) \cdot \prod_{\substack{r=1..W,\\ r\neq x}} p\left(R_{r,y} \mid d_{x,y}\right)}{\prod_{l=1..W} p\left(L_{l,y}\right) \cdot \prod_{r=1..W} p\left(R_{r,y}\right)} \cdot p\left(\left(L_{x+d_{x,y},y}, R_{x,y}\right) \mid d_{x,y}\right) \cdot p\left(d_{x,y}\right) \tag{6}$$

Also, with the exception for the mentioned case (5), the probability distributions related to $p\left(L_{l,y} \mid d_{x,y}\right)$ and $p\left(R_{r,y} \mid d_{x,y}\right)$ are independent from $d_{x,y}$ and thus can be reduced with the denominator:

$$p_{x,y,d} = \frac{1}{p\left(L_{x+d_{x,y},y}\right) \cdot p\left(R_{x,y}\right)} \cdot p\left(\left(L_{x+d_{x,y},y}, R_{x,y}\right) \mid d_{x,y}\right) \cdot p\left(d_{x,y}\right) \tag{7}$$

It can be further seen that term $p\left(L_{x+d_{x,y},y}\right)$ is probability distribution of luminance values in the left view (which is independent of particular pixel position) and can be expressed as $p\left(L_{x,y}\right)$. We finally get:

$$p_{x,y,d} = \frac{1}{p\left(L_{x,y}\right) \cdot p\left(R_{x,y}\right)} \cdot p\left(\left(L_{x+d_{x,y},y}, R_{x,y}\right) \mid d_{x,y}\right) \cdot p\left(d_{x,y}\right) \tag{8}$$

Further in the paper this formula will be used to propose a novel depth estimation method with the use of Maximum A posteriori Probability (MAP) rule (3) but, in the meanwhile, we can notice it can be simplified in order to attain classical SSD and SAD pixel similarity metrics that are commonly used in depth estimation algorithms. The term $p\left(\left(L_{x+d_{x,y},y}, R_{x,y}\right) \mid d_{x,y}\right)$ is a joint probability that luminance value $L_{x+d_{x,y},y}$ of pixel in the left view and luminance value $R_{x,y}$ of pixel in the right view will occur, on the condition that those pixels are corresponding to each other under depth $d_{x,y}$. Again, according to Bayes rule, it can be expressed as $p\left(R_{x,y}\right) \cdot p\left(L_{x+d_{x,y},y} \mid \left(R_{x,y}, d_{x,y}\right)\right)$. Therefore, the term $p\left(R_{x,y}\right)$ simplifies with the term in the denominator of (8):

$$p_{x,y,d} = \frac{p\left(d_{x,y}\right)}{p\left(L_{x,y}\right)} \cdot p\left(L_{x+d_{x,y},y} \mid \left(R_{x,y}, d_{x,y}\right)\right) \tag{9}$$

Let‘s assume the following:
A1. The presence of additive noise, the same in both of the views (in particular, with equal standard deviation σ).
A2. Lambertian model of reflectance in the scene, which means that the observed light intensity of given point in the scene is independent from the angle of viewing, and thus is equal amongst the views.

A3. Color correspondence between the views, which means that color profiles of the cameras are compatible, so that given light intensity is represented as the same luminance value μ among the views (in the consideration, for given pair of corresponding pixels $L_{l,y}$ in the left view and $R_{r,y}$ in the right view).

If we consider gaussian distribution of the noise, with mean value μ and standard deviation σ, then $L_{l,y} \sim Gaussian_{(\mu,\sigma)}$, and $R_{r,y} \sim Gaussian_{(\mu,\sigma)}$. In the term $p\left(L_{x+d_{x,y},y} \mid R_{x,y}, d_{x,y}\right)$ random variable $R_{x,y}$ is assumed to be a posteriori observation with given, specific value (also as $d_{x,y}$ is considered conditionally too), therefore $\mu = R_{x,y}$. Thus, the pixels are assumed to correspond to each other and thus both random variables have the same expected value $\mu_{x,y}$. Moreover, the difference in luminance between $L_{x+d_{x,y},y}$ and $R_{x,y}$ results only from the probability distribution $Gaussian_{(R_{x,y},\sigma)}\left(L_{x+d_{x,y},y}\right)$ of the noise, where both $R_{x,y}$ and $L_{x+d_{x,y},y}$ are our a posteriori observations:

$$p\left(L_{x+d_{x,y},y} \mid R_{x,y}, d_{x,y}\right) = \frac{1}{\sigma\sqrt{2\pi}} \cdot \exp\left(-\frac{1}{2\sigma^2}\left(L_{x+d_{x,y},y} - R_{x,y}\right)^2\right) \quad (10)$$

therefore we get:

$$p_{x,y,d} = \frac{p\left(d_{x,y}\right)}{p\left(L_{x,y}\right)} \frac{1}{\sigma\sqrt{2\pi}} \cdot \exp\left(-\frac{1}{2\sigma^2}\left(L_{x+d_{x,y},y} - R_{x,y}\right)^2\right) \quad (11)$$

We are looking for depth with Maximum A posteriori Probability and thus we search for the best matching depth d^* which has the highest (maximal) probability $p_{x,y,d}$. It is equivalent to finding d with minimal $-\log\left(p_{x,y,d}\right)$. After natural logarithm on both sides of the equation (11) is taken we get:

$$-\log\left(p_{x,y,d}\right) = -\log\left(p\left(d_{x,y}\right)\right) + \log\left(p\left(L_{x,y}\right)\right) + \log\left(\sigma\sqrt{2\pi}\right) + \frac{1}{2\sigma^2}\left(L_{x+d_{x,y},y} - R_{x,y}\right)^2 \quad (12)$$

It can be noticed that if all terms except the last one (on the right) are omitted, the equation (12) simplifies to SSD formula for pixel similarity metric:

$$-\log\left(p_{x,y,d}\right) = \left(L_{x+d_{x,y},y} - R_{x,y}\right)^2 \quad (13)$$

The omitted terms $p\left(d_{x,y}\right)$ and $p\left(L_{x,y}\right)$, $\log\left(\sigma\sqrt{2\pi}\right)$ and $\frac{1}{2\sigma^2}$ correspond to: probability distribution of depth values, probability distribution of luminance values in the left view, constant offset and constant scaling factor, respectively. Such omission could be justified if all of those terms were constants which would be true if we add two more assumptions to our considerations:

A4. Distribution of $p\left(d_{x,y}\right)$ is uniform.

A5. Distribution of $p\left(L_{x,y}\right)$ is uniform.

Analogous reasoning can be performed for the presence of Laplace distribution of the noise $L_{l,y}, R_{r,y} \sim Laplace_{(\mu,b)}$. In such a case we get:

$$-\log\left(p_{x,y,d}\right) = -\log\left(p\left(d_{x,y}\right)\right) + \log\left(p\left(L_{x,y}\right)\right) + \log\left(2b\right) + \frac{1}{b}\left|L_{x+d_{x,y},y} - R_{x,y}\right| \quad (14)$$

Here, we can see that if all terms except the last one (on the right) are omitted, the equation (14) simplifies to SAD formula for pixel similarity metric:

$$-\log\left(p_{x,y,d}\right) = \left|L_{x+d_{x,y},y} - R_{x,y}\right| \quad (15)$$

Again, the omitted terms, $p\left(d_{x,y}\right)$, $p\left(L_{x,y}\right)$, $\log\left(2b\right)$ and $\frac{1}{b}$ correspond to: probability distribution of depth values, probability distribution of luminance values in the left view, constant offset and constant scaling factor, respectively. Such omission could be justified if all of those terms were constants which would be true if both of the mentioned probability distributions (A4 and A5) were uniform.

We can thus conclude, that usage of SSD (Sum of Squared Differences) / SAD (Sum of Absolute Differences) metric is optimal (from Maximum A posteriori Probability point of view) for the case of presence of additive (assumption A1) gaussian (SSD) / Laplace (SAD) noise, independent between the views, Lambertian model of reflectance (A2), color correspondence (A3), uniformity of distributions of possible disparities (A4) and luminance (A5) values.

For the sake of brevity we omit verification of these assumptions, which can be found in [3][17]. Here we only conclude that in most of the cases, the assumptions are not true for the tested sequence data set. The probability distributions of luminance (A4) and depth (A5) values for the tested sequences are clearly not uniform. For an another example, in Fig. 3.1 (left) we can see that measured distribution of noise in exemplary *Poznan Carpark* sequence [16] is similar to gaussian but it is slightly skewed in such a way, that the maximum of the distribution is at position of about 0.4. In Fig. 3.1 (right) we can see that also there is evidence that either or both assumptions A2 or A3 are not true, because the relation between luminance vales of pixels corresponding in two views (denoted X and Y) is not linear. In the tested set ([3], Table 1) only synthetic *Undo Dancer* [23] sequence conforms the assumptions.

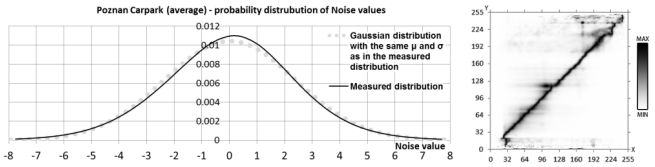


Fig. 1. Measured probability distribution of noise values [3], averaged over all views (left) and 2-dimensional histogram of luminance values (in logarithmic gray-level scale) of corresponding pixels in the views $X = 4$ and $Y = 3$ (right) of *Poznan Carpark* sequence [16].

As mentioned above, the simplifications leading to simplification of (8) to SAD or SSD are not justified in the case of the tested set. Therefore we propose to use formula (8). We express it in a logarithmic scale in decibels (thus 10 scaling factor) which is a common trick used in formulation of energy cost and probability functions for optimization algorithms [3]:

$$DataCost_{x,y}\left(d_{x,y}\right) = -10 \cdot \log\left(p_{x,y,d}\right) \tag{16}$$

For an practical application, all of the terms of probability in (8) have to been modeled. Therefore, we have empirically measured distributions of $p\left(L_{x,y}\right)$

and $p(R_{x,y})$ as histograms of the input pictures, as those terms do not depend on pixel correspondence related to depth $d_{x,y}$. On the other hand, probability distribution of depth $p(d_(x,y))$, and probability of corresponding luminance values in the left and the right view $p\left(\left(L_{x+d_{x,y},y}, R_{x,y}\right) \mid d_{x,y}\right)$ depend on depth $d_{x,y}$. Having a ground truth depth map for a given scene, both of those terms can be directly modeled. $p(d_{x,y})$, which is probability distribution of depth $d_{x,y}$, has been estimated as a histogram of the given ground truth depth map. $p\left(\left(L_{x+d_{x,y},y}, R_{x,y}\right) \mid d_{x,y}\right)$ is a 2-dimensional probability distribution that has been estimated as a 2-dimensional histogram of luminance values $L_{x+d_{x,y},y}$ and $R_{x,y}$ of pixel pairs, which are known to correspond to each other, basing on given depth value $d_{x,y}$ from the ground truth depth map (example of such histogram is presented in Fig. 3.1 right).

3.2 TransitionCost component

Similarly to previous section, we propose a probabilistic model for $TransitionCost$. We assume that $TransitionCost_{p,q}(d_p, d_q)$ can be modeled basing on probability that given two neighboring pixels p and q have depths d_p and d_q respectively. Just like before, we use logarithmic decibel scale, so that it could be used directly inside of state-of-the-art depth estimation algorithms [14]:

$$TransitionCost_{p,q}(d_p, d_q) = -10 \cdot \log\left(p_{2D}(d_p, d_q)\right) \tag{17}$$

For real data $p_{2D}(d_p, d_q)$ can be measured as 2-dimensional histogram of depth value pairs d_p and d_q of neighboring pixels p and q. In our work, this has been performed over all frames of all used test sequences and all views for which ground truth depth data is available in the test set. The exemplary graphs with the measured data are shown in Fig. 3.2 in the left column. It can be noticed that the maximum of the curves lay approximately along the diagonal but also there are strong bands on both sides.

Because often TransitionCost is expressed as a function of a single argument $|d_p - d_q|$, instead of two independent arguments, it is interesting to also see whether such formulation is justified. In order to do that, apart from figures presenting $p_{2D}(d_p, d_q)$ as 2-dimensional plots, also 1 dimensional plots of probability of given disparity difference $d_p - d_q$, $p_{1D}(d_p - d_q)$, have been visualized see Fig 3.2 in the right column. The plots are firstly falling approximately linearly and then they reach plateau until the limits of the histogram plot. Such curves resemble the shapes of linear model and truncated-linear model of $TransitionCost$. Therefore we can conclude that those classical models (linear and truncated-linear) may be adequate for the case, when the $TransitionCost$ express probability in a logarithmic scale (in which $TransitionCost$ has been depicted in figures).What is important in case of each sequence, $TransitionCost$ has different scale (slope of the curve). Without the knowledge coming from empirical analysis of the $TransitionCost$, performed likewise as in the work, this scale would have to be calibrated manually of experimentally (e.g. with use of Smoothing Coefficient in DERS [14]). This is an important advantage of the proposal.

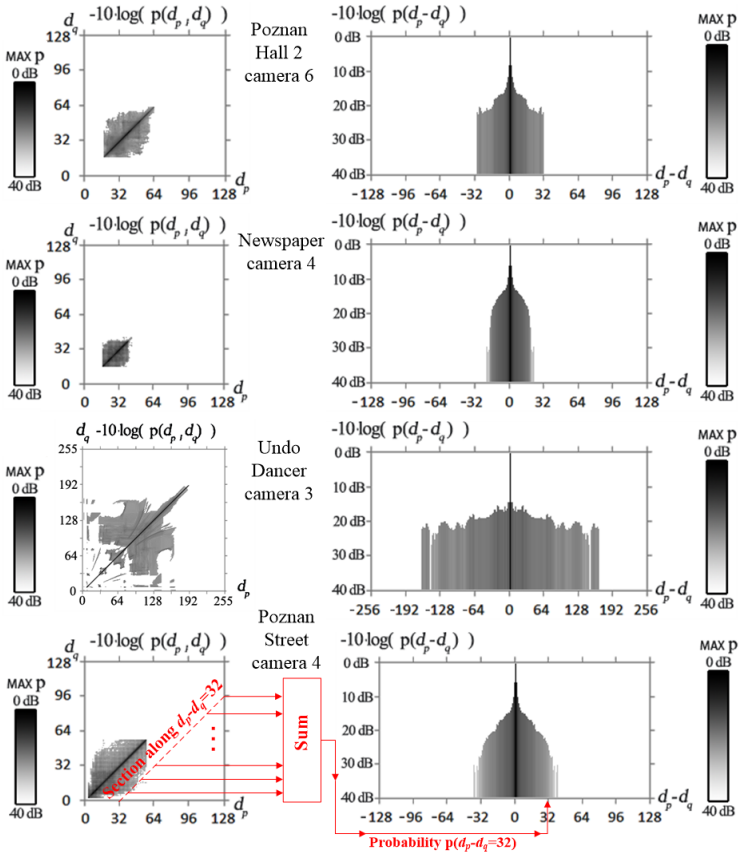
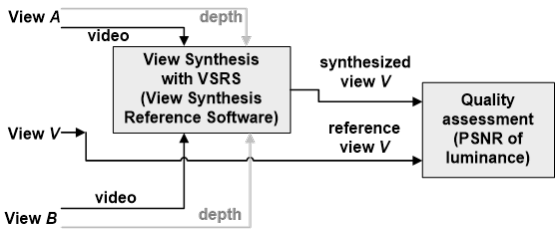


Fig. 2. Distributions of probability that neighboring pixels p and q in the ground truth depth map have depth (disparity) values d_p and d_q. Measured as 2-dimensional histograms $p_{2D}\left(d_p, d_q\right)$ (on the left) and 1-dimensional histograms $p_{1D}\left(d_p - d_q\right)$ (on the right). Exemplary calculation of $p_{1D}\left(d_p - d_q = 32\right)$ from $p_{2D}\left(d_p, d_q\right)$ has been shown in red. All plots are in logarithmic scale



Sequence Name	View ids.		
	A	B	V
Poznan Street [16]	3	5	4
Poznan Carpark[16]	4	5	4
Poznan Hall 2 [16]	5	7	6
Lovebird1[20]	3	5	4
Newspaper [21]	4	6	5
Balloons [22]	3	5	4
Kendo [22]	3	5	4

Fig. 3. Depth map quality assessment procedure used in the work.

4 Experimental results

In the previous subsections 3.1 and 3.2 we have derived probabilistic models for *DataCost* and $TransitionCost$. Those two models have been used together as a complete model for Fitness function (1) in experimental assessment described below. The tests have been performed following the ISO/IEC MPEG methodology, constituted as a part of 3D framework [18]. It employs view synthesis for evaluation of quality of depth maps, which can be used to evaluate depth estimation algorithm itself. During the evaluation, three views of each test sequence are explicitly considered A, B and V (Fig. 3.2). First, for view A and view B depth maps are estimated with use of some side views (e.g. views A-1, A and A+1 for depth estimation of view A). The estimated depths of view A and view B, along with their original images, are used to synthesize a virtual view in position of middle view V. The original image of view V is used as a reference for PSNR-based quality measurement, which provides indirect evaluation the depth map estimation algorithm used. Therefore, the quality of the depth is assessed indirectly by evaluation of quality of synthesized view.

For view synthesis we have used MPEG View Synthesis Reference Software (VSRS) [19]. As a reference depth estimation algorithm we have employed MPEG Depth Estimation Reference Software (DERS) version 5.1 [14]. The proposed *DataCost* and $TransitionCost$ models have been implemented into DERS by replacing the original Fitness function. The original (unmodified) DERS algorithm is a supervised algorithm in a sense, that special control parameter Smoothing Coefficient has to be given. Therefore, a wide range of Smoothing Coefficient has been tested. For the sake of brevity, the best and the worst performing settings for each sequence has been identified.

The overall results are presented in Table 1. It can be seen that the results of DERS with the proposed probabilistic model are very similar to the best case of the original (unmodified) DERS in most of the cases and are very little better in some cases. In average over the tested sequences, the proposed method provides about 0.08dB gain over the best identified case generated with the original, unmodified DERS (with manually crafted Smoothing Coefficient per sequence) and about 2.79dB gain over the worst case generated by DERS.

The most important thing to notice is that the proposed depth estimation technique does not require any manual settings (usage of such depth estimation is thus unsupervised). The employed Fitness function model, based on Maximum A Posteriori rule is inhered from the knowledge coming from analysis of the $TransitionCost$. Therefore, the proposed depth map estimation method has been tested only once in one configuration.

5 Conclusions

A derivation of *DataCost* based on Maximum A posteriori Probability (MAP) rule has been presented. It has been shown that some of the conditions needed for simplification to SSD or SAD forms are not met and basing on that an improved depth estimation technique has been proposed. A method for estimation

Table 1. Gains attained with joint usage of the proposed *DataCost* and *TransitionCost* models, related to the best and the worst results attained by the original (unmodified) DERS, depending on Smoothing Coefficient parameter setting

Sequence name	PSNR [dB] virtual view versus the original view		
	DERS - the worst[1]	DERS - the best[2]	Proposed[3]
Poznan Street [16]	27.56	31.98	32.02
Poznan Carpark [16]	29.05	30.71	30.95
Poznan Hall 2 [16]	32.17	32.85	32.81
Lovebird1 [20]	27.09	29.80	29.83
Newspaper [21]	27.86	31.91	31.95
Balloons [22]	29.95	32.94	32.98
Kendo [22]	33.02	35.46	35.69
Average	29.53	32.24	32.32
Average gain of the proposal	+2.79	+0.08	-

[1] Original (unmodified) DERS - the worst setting of Smoothing Coefficient.
[2] Original (unmodified) DERS - the best setting of Smoothing Coefficient.
[3] Proposed probability-based model implemented in DERS.

of parameters of this model has been shown on an example of the test sequences. Next, a probabilistic model for *TransitionCost* has been proposed also with a method for estimation of parameters of this model. In the end experimental verification has been conducted. The attained results show average gain of about 0.08dB to 2.8dB, calculated with respect to PSNR of virtual views, synthesized with use of depth maps generated with the proposed method, over the reference. As a reference, original unmodified MPEG Depth Estimation Reference Software has been used with manual calibration of Smoothing Coefficient per sequence. For the case of selection of the worst checked Smoothing Coefficient value for the original DERS, the gain is about 2.8dB of PSNR, averaged over all of the tested sequences. For the case of selection of the best found Smoothing Coefficient in original DERS software, the average gain is only about 0.08dB of PSNR, but it can be noted that the proposed technique attained that without manual calibration of such coefficient. This constitutes one of the biggest advantages of the proposed depth estimation method it does not require arbitrary manual calibration of parameters like Smoothing Coefficient. All required model parameters can be algorithmically estimated.

References

1. K. Müller, P. Merkle, T. Wiegand, "3-D video representation using depth maps", Proceedings of the IEEE, Vol. 99, No. 4, pages 643-656, April 2011.
2. M. Domański, K. Klimaszewski, J. Konieczny, M. Kurc, A. Łuczak, O. Stankiewicz, K. Wegner, "An experimental free-view television system", Image Processing and Communications Challenges, R. Chorąś, A. Zabłudowski (eds.), Academy Publishing House EXIT Warsaw, pages 169-176, 2009.

3. O. Stankiewicz, "Stereoscopic depth map estimation and coding techniques for multiview video systems", Ph.D. dissertaation, Faculty of Electronics and Telecommunications, Poznan University of Technology, Poznań, 2014.
4. M.F. Tappen, et Al. "Comparison of graph cuts with belief propagation for stereo, using identical MRF parameters", IEEE Int. Conf. on Computer Vision, 2003.
5. T. Kanade, M. Okutomi, "A Stereo Matching Algorithm with an Adaptive Window: Theory and Experiment", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 16, No. 9, pages 920-932, Sept. 1994.
6. Y. Boykov, et Al., "A variable window approach to early vision", IEEE Trans. Pattern Anal. Mach. Intell., Vol. 20, No. 12, pages 12831294, Dec. 1998.
7. Jian Sun, et Al. "Stereo Matching Using Belief Propagation", IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 25, Issue: 7, pp.787-800, 2003.
8. R. Zabih, J. Woodfill, "Non-Parametric Local Transforms for Computing Visual Correspondence", Proceedings of European Conference on Computer Vision, 1994.
9. K. Wegner, O. Stankiewicz, "Similiarity measures for depth estimation", 3DTV-Conference 2009 The True Vision Capture, Potsdam, Germany, 4-6 May 2009.
10. L. Cheng, T. Caelli, "Bayesian Stereo Matching", Proc. Conf. Computer Vision and Pattern Recognition Workshop, pages 192-192, 2004.
11. Li Zhang, S.M. Seitz, "Estimating Optimal Parameters for MRF Stereo from a Single Image Pair", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.:29 , Issue: 2, pages 331 - 342, 2007.
12. P. Viola, W. Wells, "Alignment by Maximization of Mutual Information", 1995. Proceedings of Fifth International Conference on Computer Vision, pages 16-23, Cambridge, MA , USA, 20-23 Jun 1995.
13. S. Geman, G. Geman "Stochastic Relaxation, Gibbs Distribuition and the Bayesian Restoration of Images", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 6, pages 721-741, 1984.
14. M. Tanimoto, T. Fujii, K. Suzuki, "Video Depth Estimation Reference Software (DERS) with Image Segmentation and Block Matching", ISO/IEC MPEG M16092, Lausanne, Switzerland, Feb. 2009.
15. D.M. Greig, et Al. "Exact maximum a posteriori estimation for binary images", Journal of the Royal Statistical Society Series B, 51, p. 271279, 1989.
16. J. Stankowski, K. Klimaszewski, O. Stankiewicz, K. Wegner, M. Domański, "Preprocessing methods used for Poznan 3D/FTV test sequences", ISO/IEC JTC1/SC29/WG11 MPEG 2010/M17174, m17174, Kyoto, Japan, January 2010.
17. O. Stankiewicz, M. Domański, K. Wegner, "Analysis of noise in multi-camera systems", 3DTV Conference 2014, Budapest, Hungary, 2-4 July 2014.
18. "Overview of 3D video coding", ISO/IEC JTC1/SC29/WG11, Doc.N9784, Archamps, France, May 2008.
19. "View synthesis algorithm in view synthesis reference software 3.0 (VSRS3.0)", ISO/IEC JTC1/SC29/WG11 Doc. M16090, Feb.2009.
20. Gi-M. Um, G. Ban, Yo-S. Ho, et Al. "Video Test Material of Outdoor Scene", ISO/IEC JTC1/SC29/WG11, MPEG/ M15371, Archamps, France, April 2008.
21. Yo-Sung Ho, E.-K. Lee, C. Lee, "Video Test Sequence and Camera Parameters", ISO/IEC MPEG M15419, Archamps, France, April 2008.
22. M. Tanimoto, T. Fujii, N. Fukushima, "1D Parallel Test Sequences for MPEG-FTV", MPEG M15378, Archamps, France, April 2008.
23. D. Rusanovskyy, P. Aflaki, M. M. Hannuksela, "Undo Dancer 3DV sequence for purposes of 3DV standardization", ISO/IEC JTC1/SC29/WG11, Doc. M20028, Geneva, Switzerland, March 2011.