TRANSFORM DOMAIN SIMILARITY MEASURES IN STEREO MATCHING

Olli Suominen, Atanas Gotchev

Tampere University of Technology
Department of Signal Processing
Korkeakoulunkatu 10, 33720, Tampere, Finland
olli.j.suominen@tut.fi, atanas.gotchev@tut.fi

Miska M. Hannuksela

Nokia Research Center
Multimedia Technologies
Visiokatu 1, 33720, Tampere, Finland
miska.hannuksela@nokia.com

ABSTRACT

A Fourier-domain representation of an image exhibits a property where a translation of the image is included in the phase term. This property extends to the Discrete Cosine Transform, where the translation is encoded in the signs of the coefficients. An interpretation of this property for use as a similarity measure in local stereo matching is explored with an efficient way of comparing the transformed blocks to generate a dense disparity map. The method is empirically demonstrated to work also with the Haar wavelet transform, which offers faster computation and improved quality. Results show that presented transform based similarity measures provide better disparity estimates than box aggregated Sum of Absolute Differences and the Census transform when using the percentage of bad pixels as the quality measure. In terms of mean squared error, they achieve better results than SAD but are marginally below Census. The simplicity of making the comparisons results in good scaling with the number of disparity estimates, making the suggested method perform better than SAD also computationally.

Index Terms **—** Stereo image processing, Discrete cosine transforms, Discrete wavelet transforms

1. INTRODUCTION

Image registration and stereo matching have some common elements, namely finding correspondences between images. Phase-only correlation in Fourier domain has been used extensively in image registration, and has received some attention for usage also in stereo matching algorithms [1, 2]. In the time scale of image registration research, a relatively recent suggestion is to utilize Discrete Cosine Transform's characteristic as a special case of DFT for a simpler registration method [3]. The purpose of this paper is to study the use of DCT sign-only correlation as a similarity measure in a local stereo matching environment. The idea of coefficient sign correlation is also extended to an integer transform that is designed to approximate DCT, and a wider class of transforms, namely Walsh-Hadamard and Haar.

Phase only correlation (POC) can be used to find the translation, rotation and scaling between two images. It has often been applied to various image registration applications. In the particular case of stereo matching, only detecting translations of image segments is required. A detailed description between POC and DCT sign only correlation can be found in [3]. In this context, it is sufficient to highlight the transform domain similarity S_T between two real valued $N \times M$ images f and g,

$$s_T(f,g) = \frac{1}{NM} \sum_{n=1}^{N} \sum_{m=1}^{M} F'_T(n,m) \cdot G'_T(n,m), \qquad (1)$$

where $F'_T(n,m)$ is the sign $(+1/-1)$ of the transform coefficient $F_T(n,m)$.

1.1. Transforms

While there is a solid theory on the properties of POC and how it applies in the special case of DCT, the concept is not in practice limited to Fourier related transforms. The approach can be easily applied to any generalized harmonic transform that produces signed coefficient values. The transforms considered in the experimental section of this study are DCT, two integer transforms designed to approximate DCT, Walsh-Hadamard and a modified Haar transform.

There is an existing method of speeding up computation of DCT coefficients for adjacent, overlapping blocks [4], which also could be used to exploit source images compressed with DCT based encoding. Also, as DCT is used in many performance oriented applications, several approximations for it have been introduced. They rely on converting the floating point coefficients of the true DCT matrix to integers while retaining the properties of an orthogonal transform, so that the computation is simplified. The effect of this is tested with the transform matrix from [5].

Another simple and relatively easy to compute transform is the Walsh-Hadamard transform. The transformation matrix consists only of ones and minus ones, and the scalar scaling of the matrix can be dropped to simplify computation even further. Although it does have a relationship with the Fourier transform, the potential link is not explored here any further. [6]

Using the zero crossings extracted via a wavelet transform for stereo matching has been suggested before, but instead of sign correlation, the similarity measure itself was based on sum of squared differences [7]. As we are not interested in reconstructing the spatial domain presentation from the transformed window, the requirement of orthogonality can be relaxed. Therefore the transform matrix can be further simplified by dropping the scaling coefficients, which together with the large number of zero coefficients in the matrix enables effective optimization [6]. This has very little effect on the results. A straightforward unrolling of the matrix multiplication yields 14 +/- operations for transforming a

vector of 8 samples. Assuming 2D transforms are done in two passes, a rough estimate for transforming an 8x8 window centered on a pixel would take 126 operations, down from the 2000 of a naive multiplication.

2. STEREO MATCHING

In this study, the area of interest is the application of transform based metrics in local stereo matching methods. The intent of stereo matching is to find correspondences between two images that are taken from the same scene, but from a slightly different viewpoint. For local stereo matching, the correspondences are considered for a spatial neighborhood of each location. In a properly rectified image, the correct match will be found on the same horizontal line, i.e. epipolar line. Therefore it is sufficient to look for correspondences in just one dimension, creating a one-dimensional cost vector with an element for each disparity that is tested. The best disparity estimate for that neighborhood is then the disparity that gives to the maximum or minimum cost, depending on the similarity measure(or cost function) used. [8]

The proposed similarity measure here is Eq. 1. The measure is computed for each blockwise comparison, comparing a reference block from the other image to a number of blocks on the epipolar line from the other image. The disparity D at coordinates x, y is therefore

$$D(x,y) = \underset{d \in [D_{min}, D_{max}]}{\arg\max} \; s_T(B_L(x,y), B_R(x+d,y)), \quad (2)$$

where $B_L(x,y)$ and $B_R(x,y)$ are equally sized blocks from left and right images centered on x, y, and D_{min} and D_{max} the limits of the disparity range being searched. While 1 is presented using Discrete cosine transforms, the transform itself can be substituted e.g. with the ones from section 1.1.

2.1. Encoding coefficient signs

In this particular application, any individual transform coefficient can have three different states: greater than, equal to or less than zero. The sign of the zero coefficients can be considered to be ambiguous, so it is defined to be +. This enables a binary representation of a single coefficient sign, i.e. 0 for -, 1 for +. While this also saves some memory space in a practical implementation, the main benefit is the computational simplicity gained. For an $N \times N$ sized transform, the relevant information is contained in only N^2 bits. Theoretically there are no constraints for the size of the window, but again from a practical point of view, not only is the block size of 8x8 both common in fast transform implementations, but also results in a 64 bit representation for a single window, which efficiently and evenly maps into one or two registers in modern CPUs. Later on this representation is called a *bit string*. This formulation leads to similar computation that has been used e.g. in conjunction with the Census transform [9].

2.2. Pairwise comparison of encoded signs

When finding matches between the images in a stereo pair, the transformed and encoded windows are compared to the candidate positions in the corresponding pair. The correlation between the encoded signs is reduced to being computed like the Hamming distance between the bit strings. This can then be computed by taking a XOR of the two strings and counting the set bits from the result. The larger the Hamming distance, the smaller the correlation of those windows.

Table 1. Number of operations required to compute one disparity estimate between blocks. Listed methods are box filtered SAD (with summed area tables) and two approaches to comparing transform based bit strings (with 16 bit memory lookup or hardware implemented population count).

	Arithm. op	Mem reads	Mem writes
Box SAD	5	6	3
SW pop. count	4	5	1
HW pop. count	2	1	1

A general solution for counting set bits that can be used on most platforms is to use a precalculated lookup table. The XORed bit string is interpreted as an integer, and is used to index the lookup table. For instance, using 16 bit segments, bits of a 64 pixel bitstring is processed with 4 memory lookups. A more specialized solution is to use hardware dependant operations directly designed for the task. Such operations are for instance POPCNT in some x86 processors, or VCNT in NEON accelerators commonly found in ARM processors. Availability of such a command greatly affects the speed of making the comparisons (Table 1).

2.3. Filtering of cost volume and disparity map

In a traditional setting, the assumption that pixels spatially close to each other tend to have similar disparities is exploited by combining all of the pixel-wise cost vectors, creating a 3-dimensional cost volume. Each disparity estimate level corresponds to a XY-slice in that volume, which can be then be spatially filtered to enforce the assumption. [8] Doing the cost volume filtering for transform based methods is in some respect redundant as the neighborhood is already considered in the original windowing, but there is still an improvement in quality to be gained. The performance implications are quite severe, though, as each disparity estimate requires a separate application of the filter of choice. The cost of any operations done on slices of the cost volume is multiplied by the amount of disparity estimates, which is why more advanced and computationally expensive (e.g. edge-aware) cost aggregation methods are not considered here.

After the disparity estimate is created from the cost volume, an extremely simple post processing method that is of interest in this context is rank order filtering. The outlier matches among the correct matches can be interpreted as noise in the signal. The noise can be treated as statistically independent from the true signal, as there are no strong constraints why the erroneous best matches would be close to the correct one. Linear filters such as the Gaussian or moving average handle that kind of noise poorly, so rank order filtering is preferrable. Especially without a separate aggregation step, a single iteration of median filtering on the result offers a major gain in quality.

3. EXPERIMENTAL RESULTS

The similarity metrics used as a reference point is the commonly used Sum of Absolute Differences (SAD) and the Census transform. The quality of disparity estimates is evaluated with the percentage of pixels that do not match the ground truth (BAD) and the Mean Squared Error of the estimate against the ground truth. The data set for these experiments is 21 pre-rectified1/3 sized Middlebury stereo pairs [10]. Occluded areas are extracted from the

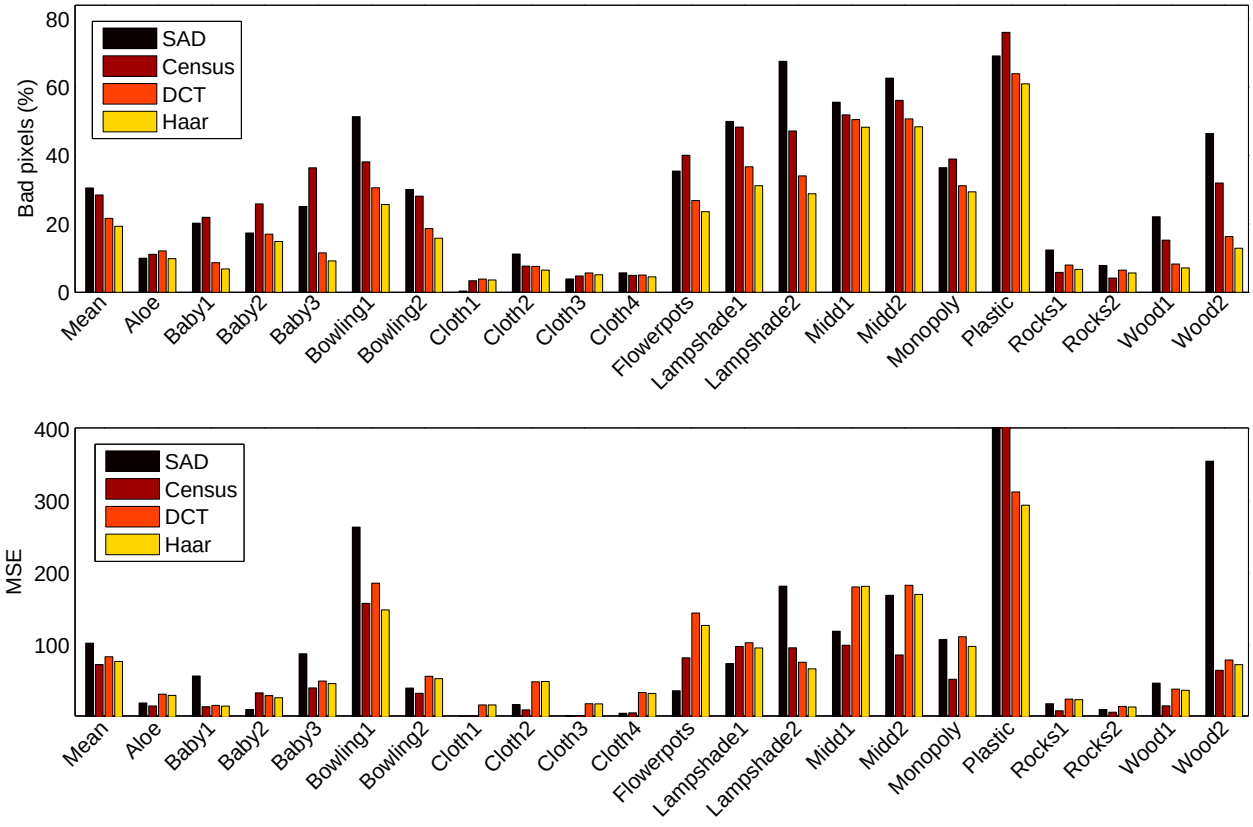

Figure 1. Percentage of bad pixels and MSE of estimated disparity map for each image in the Middlebury 2006 data set using SAD, Census, DCT and Haar based similarity measures. The bars that are cut off for Plastic go up to approx. 600. All disparity maps have been 5x5 median filtered.

ground truth and are excluded from the numerical quality estimates, as any kind of occlusion filling is not being performed or proposed. For the reference metric SAD, the filter radius of cost volume aggregation is always selected as the one providing best results in terms of that specific quality measure from the range $[0 - 11]$, while while Census is used with the same 8x8 as the proposed transforms.

As shown in Table 2, all the proposed transforms perform on similar levels. This is also true for the majority of individual comparisons. The values are the average over the whole data set, which includes scenes of varying difficulty, which makes the values relatively high. In light of this comparison, the rest of the results are presented only for DCT and Haar for clarity. DCT was selected due to its widespead usage in image processing, and Haar due to it displaying the best performance. Figure 1 shows percentage of bad pixels and MSE for each individual stereo pair in the data set. The transform based methods give disparity estimates with less bad pixels for the majority of the pairs, and do not fall very far behind even in the cases where SAD performs better. In terms of MSE there is some more variation between stereo pairs, with the averages over the whole data still favoring the proposed methods. Notable is though that Census is ahead when measured in MSE.

Figure 3 demonstrates the effect of an aggregation step on the cost volume. In practice, SAD requires an aggregation step, but for transform based methods it is optional. Figure 2 shows examples of disparity estimates for the Bowling1 stereo pair using SAD, DCT and Haar. The estimates clearly show the different types of artifacts that originate from the different similarity measures. As shown in Figure 4, it is clear that the minor effort of a single pass of median filtering helps with the type of noise, reducing bad pixels on average by up to 5 percentage points.

Table 2. Average quality metrics over the whole dataset for each of the transforms experimented with. I-DCT is the integer approximation of DCT [5]

	SAD	Census	DCT	I-DCT	Walsh	Haar
BAD	30.5%	28.5%	21.6%	21.1%	20.2%	19.3%
MSE	102.9	73.1	84.0	82.7	77.2	77.4

4. CONCLUSIONS AND EVALUATION

In most cases, the proposed transform based metrics offer better quality in comparison to SAD with box filter based aggregation. They are also notably invariant to differences in exposure between the images in the stereo pair, as it only influences one component of the transform coefficients. Another benefit against SAD is a better computational scalability with the number of disparity measures. This is due to the more compact presentation of the required information, which better utilizes the memory architecture and ca-

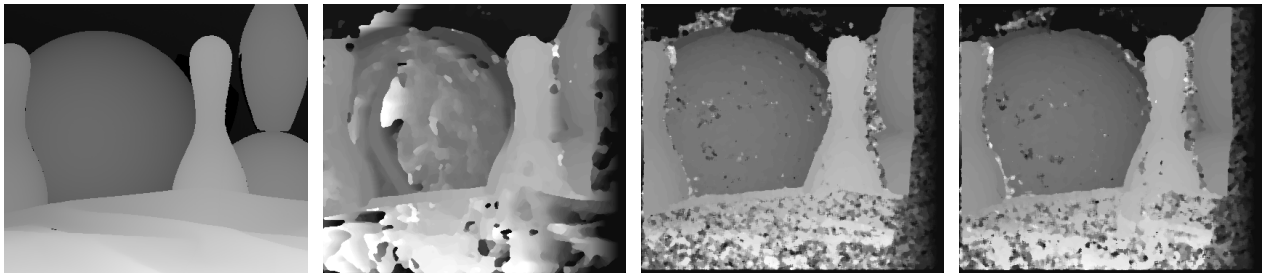

Figure 2. Disparity maps of Bowling1 from left to right: Ground truth, estimates with SAD with box filter, DCT and Haar. All maps have been 5x5 median filtered. The Bowling1 scene illustrates well the types of "noise" characteristic originating from the methods. SAD suffers from large smooth patches of wrong disparities while transform based methods experience a kind of salt&pepper type of noise.

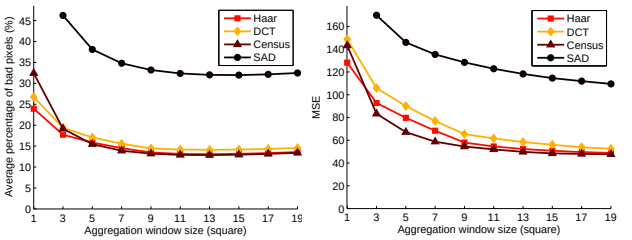

Figure 3. Average percentage of bad pixels and MSE for the Middlebury 2006 data set after box filtering (i.e. aggregating) the cost volume with different block sizes. Size 1x1 corresponds to no filtering.

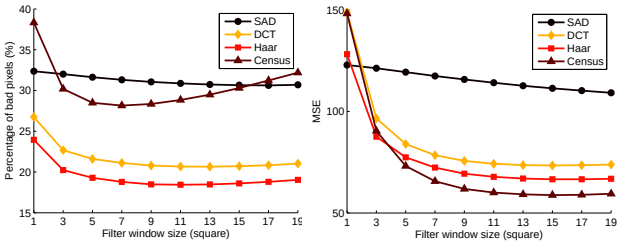

Figure 4. The effect of median filtering with different sized windows on the quality of the disparity estimate.

pabilities of modern processors. Quality can be greatly improved with simple post processing using a median filter on the disparity estimate. The Census transform also exhibits these properties, but despite being slightly above in terms of MSE, is still clearly behind on bad pixels. In total, Census seems to be wrong more often than the proposed methods, but is somewhat less wrong. It depends on the application and further post processing which one is the preferred alternative.

A cost volume created with the transform domain metrics does benefit from a separate, additional aggregation step. However, it is not required to the achieve a relatively good quality in comparison to SAD based matching, and as it would more than double the processing workload, it may not be worth the effort in cases where speed is essential. If such a compromise is made, it could also be justifiable to use Census, as it converges to similar quality levels with the extra aggregation step and is slightly faster to compute.

As far as quality and speed is considered, the Haar based matching is clearly a better choice for most applications. However, DCT is frequently used in other image processing tasks. It may allow for synergy benefits between stereo matching and other algorithms that are also transforming some or all of the same windows. An integer based approximation of DCT can work for this purpose even better than the original.

5. REFERENCES

[1] Mohammad Abdul Muquit, Takuma Shibahara, and Takafumi Aoki, "A high-accuracy passive 3d measurement system using phase-based image matching," *IEICE Trans.Fundam.Electron.Commun.Comput.Sci.*, vol. E89-A, no. 3, pp. 686–697, mar 2006.

[2] D. V. Papadimitriou and T. J. Dennis, "Stereo disparity analysis using phase correlation," *Electronics Letters*, vol. 30, no. 18, pp. 1475–1477, 1994.

[3] I. Ito and H. Kiya, "Dct sign-only correlation with application to image matching and the relationship with phase-only correlation," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, 2007, vol. 1, pp. I–1237–I–1240.

[4] V. Kober, "Fast algorithms for the computation of sliding discrete sinusoidal transforms," *Signal Processing, IEEE Transactions on*, vol. 52, no. 6, pp. 1704–1710, 2004.

[5] Honggang Qi, Wen Gao, Siwei Ma, and Debin Zhao, "Adaptive block-size transform based on extended integer 8x8/4x4 transforms for h.264/avc," in *Image Processing, 2006 IEEE International Conference on*, 2006, pp. 1341–1344.

[6] Agaian Sos, Sarukhanyan Hakob, Egiazarian Karen, and Astola Jaakko, *Hadamard Transforms*, SPIE Press, 1 edition, 2011.

[7] S. Mallat, "Zero-crossings of a wavelet transform," *Information Theory, IEEE Transactions on*, vol. 37, no. 4, pp. 1019–1033, 1991.

[8] Daniel Scharstein and Richard Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *International Journal of Computer Vision*, vol. 47, no. 1, pp. 7–42, 2002, 10.1023/A:1014573219977.

[9] Martin Humenberger, Christian Zinner, Michael Weber, Wilfried Kubinger, and Markus Vincze, "A fast stereo matching algorithm suitable for embedded real-time systems," *Computer Vision and Image Understanding*, vol. 114, no. 11, pp. 1180–1202, 11 2010.

[10] H. Hirschmuller and D. Scharstein, "Evaluation of cost functions for stereo matching," in *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, 2007, pp. 1–8.