When Is a Confidence Measure Good Enough?

Patricia Márquez-Valle[1], Debora Gil[1], Aura Hernàndez-Sabaté[1], and Daniel Kondermann[2]

[1] Computer Vision Center, Universitat Autònoma de Barcelona
{pmarquez,debora,aura}@cvc.uab.cat
[2] Heidelberg Collaboratory for Image Processing at the IWR, University of Heidelberg
daniel.kondermann@iwr.uni-heidelberg.de

Abstract. Confidence estimation has recently become a hot topic in image processing and computer vision. Yet, several definitions exist of the term "confidence" which are sometimes used interchangeably. This is a position paper, in which we aim to give an overview on existing definitions, thereby clarifying the meaning of the used terms to facilitate further research in this field. Based on these clarifications, we develop a theory to compare confidence measures with respect to their quality.

Keywords: Optical flow, confidence measure, performance evaluation.

1 Introduction

The aim of confidence estimation in image processing can be intuitively described: Given the input as well as the output of a given algorithm (*e.g.* an image pair and a flow field), how can we evaluate the "amount of uncertainty" of the result at each pixel location? Knowing how much we can trust these results is crucial for all applications with legal obligations and possibly lethal consequences in domains such as driver assistance and medical imaging systems.

In this paper, we focus on confidence measures in dense correspondence problems such as optical flow, image registration and stereo depth estimation. Without loss of generality, we will focus on optical flow algorithms. In optical flow, the accuracy is commonly computed by the end-point error [1]. If we could estimate the accuracy of the result at each pixel, we could use this information by either improving the model or by removing all results above a given error threshold.

Yet, in general, given a model together with the input as well as output data, it is impossible to estimate the accuracy. To illustrate this case, consider the example of a white image pair (*i.e.* all pixels have equal intensity). Given that the pairing of such two images could be achieved by any flow field, the output of any optical flow method will only depend on the prior knowledge added to the model, L^2-regularization for instance. But not being able to estimate the accuracy, does not mean that we cannot find out anything about the error in the flow field. By analyzing the input data along with the model we can try to estimate the *upper bound* of the error at each pixel location. This information will not help to improve the flow field, but it reveals the *risk* of assuming that the outcome is actually correct in such location.

M. Chen, B. Leibe, and B. Neumann (Eds.): ICVS 2013, LNCS 7963, pp. 344–353, 2013.

This paper contributes to analysis of confidence measures in three aspects. First, we clarify the definitions of the terms accuracy, confidence and error prediction; second, we group existing confidence measures using these terms and third, we propose an approach to assess how good a confidence measure is with respect to its capabilities for error bounding. Our explanations will be accompanied by various examples showing the usefulness of our approach. Please note that actual comparisons between existing confidence measures is beyond the scope of this paper.

1.1 Related Work

In their seminal work on optical flow evaluation, Barron et al. [2] emphasized the importance of confidence measures to examine optical flow methods and also carried out a first comparison. A few years later, Bainbridge-Smith and Lane [3] compared seven different confidence measures for two image sequences. These results have been first steps towards a comparison of confidence measures within a single framework. The importance of such a framework and a general roadmap for the evaluation of optical flow was recently discussed by an international group of researchers in [4]. In this section, we exemplarily review the literature explicitly dealing with confidence measures of several types.

Early papers were mainly based on local image properties. Chronologically, [5] only evaluate flow where the determinant of the local Hessian (i.e. the product of the eigenvectors of the second image derivatives) of the image is non-zero. In contrast to analyzing the eigenvalues of the image itself, [6] analyzes the eigenvalues of the energy distribution resulting from block-matching. This idea was later picked up and relabeled as surface measure in [7]. Similar methods can be found throughout local optical flow literature. All these measures either use the energy or the image curvature as indicators for confidence. In [8] the authors aim for a statistical analysis of the resulting flow vectors. For each pixel, they compute a two-dimensional distribution of flow vectors conditioned on the local image gradient. As the authors choose this condition to be a Gaussian based on the local image structure, this method basically is an anisotropic version of Lucas and Kanade [9] defined in a probabilistic framework. Therefore, as in [5] the inherent confidence measure is also based on local image curvature.

An early general attempt to define a type of confidence measures for global flow computation methods has been made by Bruhn et al. [10]. The authors argue that for global approaches one should always use the inverse of the energy at every pixel as a confidence measure, since it shows exactly how well the flow vectors fit the model. They compare their method to using the image gradient magnitude as confidence measure and validate the quality of the results based on sparsification curves. To create such curve, the flow field is systematically sparsified by a fixed percentage of flow vectors which are sorted according to their confidence. For each such threshold, the remaining average angular error is plotted.

Furthermore, the authors state in [10] that any other confidence measure with better performance should be integrated into the flow computation method. They conclude that for these reasons the inverse of the energy is the only real confidence measure for global optical flow methods. Hence, the authors understand the term confidence to be synonymous to error prediction. As discussed in Section 2, this holds true in the special

case when not only the upper bound of the flow error can be evaluated but also a lower bound. Yet, in general, confidence estimation can only estimate upper error bounds. As a consequence, they can not be inserted into the model since a low confidence does not necessarily imply an erroneous flow vector. Besides, as explained in [11], sparsification plots are not suitable for evaluating the quality of a confidence measure. The main problem is that the removal of pixels ordered by confidence not necessarily removes pixels with high errors. This results in possibly unfair comparisons between measures.

More recently, a confidence measure has been defined based on the statistics of empirically measured distributions of flow fields [12]. They are independent of the particular model used for flow estimation and require a database of motion patterns which is representative for the application at hand. The flow field is sparsified by the continuous p-values computed from the mahalanobis distances on the previously computed principal components. The bootstrap method proposed in [13] computes the variability of the computed flow with respect to a perturbation of the variational model, by excluding randomly selected pixels in the data term. Pixels with high variability are associated to model inconsistencies and, thus, discarded. Finally, in [14], the authors train a random forest to classify errors in the flow field which are larger than a given threshold. The features used in their approach are based on a combination of all confidence measures described above, including the image data on several scales as well as the output of the algorithm itself. Sparsification is based on the continuous output of the random forests.

2 Accuracy, Confidence and Prediction

2.1 Sources of Inaccuracies

Confidence measures aim at measuring the uncertainty of flow fields at each image pixel from the input data along with the model output. Inaccuracies in the output of optical flow algorithms follow from three main reasons: Model Assumptions, Multiple Global Minima and Numerical Stability of Local Minima.

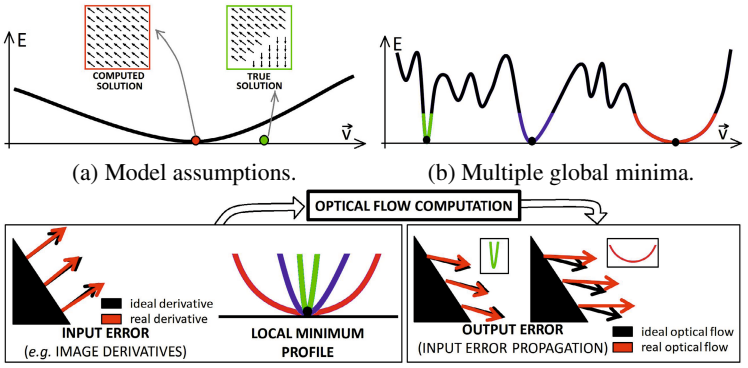
Model Assumptions. The goal of optical flow approaches is to find the vector that best matches two consecutive frames in a sequence. Given that motion vectors are at least 2D, there is not enough information in the image data alone for producing a unique solution. Consequently, optical flow algorithms need to assume some conditions on the output vector in order to compute it. Intuitively, these conditions favor a particular kind of vector field satisfying some theoretical requirements (the model assumptions) for being the final solutions to the problem. Model assumptions on optical flow can be of either analytical or probabilistic type. In the first case, the flow vector must satisfy some degree of regularity. This is usually enforced by adding the norm of the flow in some Sobolev space (usually L^2 or L^1) to a variational formulation of the optical flow problem [15]. In the second case, the flow vector should follow a given probabilistic distribution or be generated by a finite number of basic functions [15]. In any case, the restriction of possible vectors given by model assumptions might not be right for all image patches and flows. This implies that even if we have a unique stable solution, the final output might not resemble the true motion at all (see fig. 1 (a)).

Multiple Global Minima. Optical flow computation follows from the minimization of an energy functional including a data term and model assumptions. Unless simplified models are used (such as classic Horn and Schunck [16]), this functional will not be, in general, convex. Lack of convexity introduces multiple local minima that hinder the performance of gradient descent approaches based on Euler-Lagrange equations. On one hand, varying initial conditions might lead to different solutions. On the other hand, the iterative solution might get trapped in a saddle point not reaching a minimum of the energy. Multiple minima follow from non-convexity of the energy functional that is minimizing. Although theoretically convexity can be analyzed by means of the second derivative of the variational [15], in practice it is difficult to have a friendly analytical expression and some sort of heuristics should be used. Besides, a main concern is that, even if we are able to find all possible local minima, there might not be any objective criterion to decide which is the optimal solution. This uncertainty is illustrated in the plot representing an energy function shown in figure 1(b). The energy has three local minima (depicted by dots) that have equal energy and, thus, are also global.

Numerical Stability of Local Minima. The minimum of the energy modeling optical flow requires a numeric scheme in almost all cases. In this setting, it is important ensuring that any variability in the input data will not introduce a large deviation in the solution (see bottom sketches in fig.1(c)). In mathematical numerical analysis this is called error propagation or numerical stability [17]. The main concern of numerical stability is to bound errors of the output data (ε_{out}) in terms of the error of the input data (ε_{in}) by means of a constant K such that $\|\varepsilon_{out}\| < \mathrm{K} \cdot \|\varepsilon_{in}\|$ for $\|\cdot\|$ a given norm of the space the data belongs to, usually $\mathbb{R}^n$. The constant K is an intrinsic property of the algorithm and it is computed using the energy derivatives [17]. In the special case of local minima, error propagation is directly related to the flatness of the energy around each local solution. Intuitively, if the energy local profile at a minimum is flat, the number of points locally having a low energy value increases. Thus, the position of the local minimum is less accurate. On the contrary, its location accuracy increases as the profile becomes more acute. In other words, flat profiles magnify differences in initial inputs more than acute ones (see fig. 1(c)).

2.2 Confidence Measures and Error Prediction

Confidence measures can be formulated from either an analytic or a probabilistic point of view. Analytic approaches either use the energy [10,6] or the image structure (gradient magnitude [2], structure tensor [18]) as indicators of confidence. Energy-based approaches are linked to the capability of finding the energy minima and, thus, energy convexity. Whereas structure-based approaches are related to numerical stability and model assumptions. Probabilistic approaches define confidence in terms of probabilistic distributions of either flow fields itself [12] or its variability with respect perturbations in the model [13]. Probabilistic approaches are more flexible and not necessarily linked to any source of error. Furthermore, they can even be used to get a confidence fusing all previous measures [14] and, thus, relate to all three error sources. Table 1 categorizes existing confidence measures according to their formulation and source of error they mainly focus on.



(a) Model assumptions. (b) Multiple global minima.

(c) Numerical stability of local minima.

Fig. 1. Three main sources of error in optical flow algorithms: model assumptions (a), multiple global minima (b) and numerical stability of local minima, (c)

Table 1. Categorization of Confidence Measures according to Error Types

			Model Assumptions	Multiple Minima	Numerical Stability
Measure Formulation	Analytic	gradient-based [2]	✓	×	✓
		image local structure [18]	✓	×	✓
		energy-based [10]	×	✓	×
	Probabilistic	bootstrap [13]	✓	×	✓
		p-val [12]	×	×	×
		random forest [14]	✓	✓	✓

In an ideal case, we would expect the values of a confidence measure to be correlated to the flow endpoint error. In this case, the relation between measure and error could be estimated by means of non-linear regression. The confidence values would provide an estimation of the flow error and they could be further used for predicting it in sequences without ground truth. Unfortunately, this is not possible in the general case, given that errors either follow a random distribution or can not be estimated (as the white sequence example illustrates). A more realistic approach is to define quantities that estimate an upper bound for the flow error. This is consistent with the bounds on error propagation defined in the context of numerical stability. In order that a measure is useful for bounding errors, the plot between the measure and endpoint errors should show a monotonic tendency.

The plots in figure 2 illustrate the difference between the concepts of error prediction and error bounding. For both plots, the confidence measure (cm) is plotted in the x-axis and its corresponding End-point Error (EE) [1] in the y-axis . In the first case, shown in fig. 2 (a), there is a clear functional correlation between cm and EE. The scatter along the curve determines the quality (confidence interval around the value EE_0) of the error prediction given by the measure values. In the second case, fig. 2 (b), we do not have a

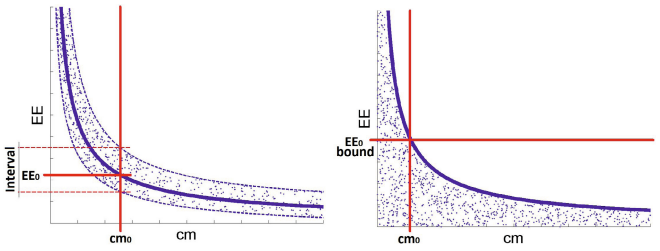


Fig. 2. Difference between Error Prediction, (a) and Error bounding, (b)

functional dependency but a sparse plot following a decreasing pattern. It follows that a given value of cm is able to provide, at most, an upper bound of the error values. In particular, we have that low values for the confidence do not necessarily imply a large error. That is, errors can take any possible value.

3 The Confidence of the Confidence

Once we have a confidence measure we still need a way to evaluate it. As discussed in the previous section, a confidence measure can, at most, provide a bound on the flow error. Therefore, we should be confident on cm in the measure it is good for bounding error. At a first intuitively approach, we would like the lowest possible error bound for all pixels.

In an ideal world, cm should give an upper bound for EE everywhere. That is, $\forall cm_0$, EE values should be bounded for all cm values above cm_0, so that the following probability is zero:

$$\forall cm_0, \exists EE_0 \text{ s.t. } P(cm > cm_0, EE > EE_0) = 0 \tag{1}$$

In practice, there is a percentage of points with an error that can not be bounded by the measure:

$$\exists cm_0 \text{ s.t. } \forall EE_0, P(cm > cm_0, EE > EE_0) = \alpha > 0 \tag{2}$$

We define the risk of a confidence measure as the percentage of points, α, which bound can not be determined by cm values. It should be clear that the lower the risk, the higher the power for bound prediction of cm.

The scatter plot in fig. 3(a) showing cm versus EE illustrates the concept of risk. The vertical line represents the threshold for cm at the value cm_0 and the horizontal lines several bounds on EE_0 having different risks ($risk_1 > risk_2 > risk_3$). For each EE_0 its risk is given by the percentage of points on the upper right square defined by the two lines. Given that EE_0 can take any possible value and this holds for all $cm > cm_0$, it is clear that cm can not provide a bound on the error everywhere. This is not the case for the scatters shown in fig. 3(b), where, for each cm value, equation (1) holds.

Not all $cm - EE$ scatters having the same risk have equal properties for error bound prediction. Also, for a given risk, there are several curves that could be fitted to the data

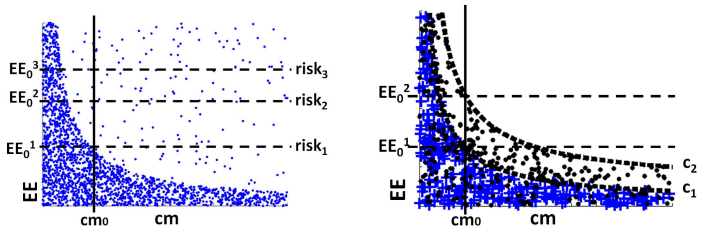


Fig. 3. Concepts involved in the quality of a confidence measure: risk in error bound prediction, (a), and optimal error bound for a given risk, (b)

for error bound setting, since the only requirement for error bounding is that curves should be monotonically decreasing. Figure 3 (b) shows two scatter plots (one in blue crosses and the other one in black dots) achieving risk zero for their envelope curves plot in dashed lines and labeled c_1 for the cross scatter and c_2 for the dot one. Yet, having both a zero risk, c_1 is better than c_2 because for any cm_0 value we have $EE_0^1 < EE_0^2$. On the other hand, note that c_1 is also a valid curve for error bounding for the second scatter plot. It provides a better error bound than c_2, but it increases the risk. This is because the curve c_1 does not enclose all points belonging to the second scatter. Those points above c_1 represent the risk.

Under the considerations above, a confidence of the confidence measure should quantify, for each risk, how good for error bounding the measure is. That is, for a given risk, each cm value should provide the lowest possible bound EE_0. We observe that the best scatter in terms of error bounding is the one having a maximum decay or, in other words, the one having a minimal area of points without risk. Given a decreasing curve fitted to the scatter, this area corresponds to the Area Under the Curve (AUC) while the risk is given by the percentage of points above the fitted curve. The trade off between the risk and EE_0 for all decreasing curves fitting the scatter data measures how good for error bounding our confidence measure is. Meanwhile, the optimal fitting curve should reach the best compromise (prone to vary across applications) between risk and AUC.

The plot showing AUC versus risk, for curves of increasing risk is our confidence of the confidence measure and we will name it RAUC. By the previous considerations, it should be clear that the steeper the RAUC is, the better the confidence measure is. The curve of minimum risk is given by the envelope to the scatter plot. It represents the worst EE_0. The convex curve enclosing the maximum number of pixels gives the lower bound for EE_0. Its risk is a measure of the maximum risk for the best EE_0 bound. The intermediate curves are computed iteratively removing a percentage of these points.

Figure 4 shows the RAUC (first on the left) and scatter plots (second to fourth) for three representative examples of a confidence measure: an ideal case ($C1$, second), expected case ($C2$, third) and worst case ($C3$, fourth). Some representative decreasing curves fitted for a given risk are also shown on each scatter. For the ideal case (C_1), the envelope of the scatter (the curve of risk 0) is the first convex curve enclosing all the points (that one with the best EE_0). It is ideal because it is able to produce a convex curve achieving zero risk. Consequently, its RAUC first point starts with the lowest

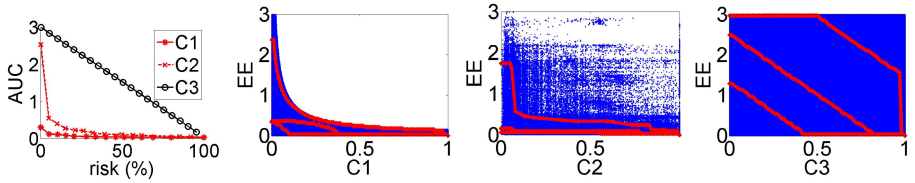


Fig. 4. Synthetic example. From left to right: the RAUC for the three different cases, ideal case ($C1$), expected case ($C2$) and worst case ($C3$).

value, and its profile is flat. The middle scatter (C_2) achieves the first convex curve at a moderate risk, which is parsed by the fitted curves. Its RAUC is worse than the case before and has positive area, so that there are still points under the curve. Finally, the right scatter (C_3) is the worse possible case, because EE is random for all cm values and thus, cm is unable to bound the error. In this case, assuming more risk, does not imply lower values for EE. Consequently, the RAUC has a linear decreasing pattern. Any confidence measure should have a RAUC under this line.

3.1 Benchmark Examples

In order to show the main concepts developed in this section, we show some examples over benchmark sequences selected from two different databases: Middelbury [1] and Sintel [19]. Optical flow was computed using the local-global approach with warping implemented by [20]. Concerning confidence measures, we chose two representative examples of each formulation category. The structured based measure defined in [11] and the energy based [10] for analytic formulations and p-val [12] and bootstrap [13] for probabilistic approaches. They are denoted as C_k, C_e, C_s and C_b respectively. We also plot C_u as a uniform confidence measure (as seen in fig.4), that is, the worst case.

For each sequence, we show the $cm - EE$ scatter plot with some of the fitted confidence curves in red and the RAUC plot for all confidence measures, as well as, the baseline worst performance C_u. Figure 5 shows results for the Middlebury sequences hydrangea and grove3 and for the Sintel sequences cave2 and sleeping1. In general, C_s scatters resemble a uniform distribution as the synthetic one shown in fig.4. This is reflected by the RAUC curves which are among the worse for all cases and are, indeed, the worst ones for sleeping1, grove3 and hydrangea. Concerning the remaining $cm's$, they show a monotonic decreasing tendency which risk and bounding capabilities vary across sequences. There is not a clear best performer for all cases due to the fact that, as discussed in Sec.2, each measure focus on a different source of error.

4 Conclusions

Confidence estimation is of prime importance for decision support systems and quite a lot of research has been recently done. Yet, there is little consensus about the meaning of some usual terms and the best way to assess the quality of a confidence measure. In this paper, we have introduced a setting for categorizing confidence measures in terms of accuracy and capabilities for error bound prediction. We have also developed a validation

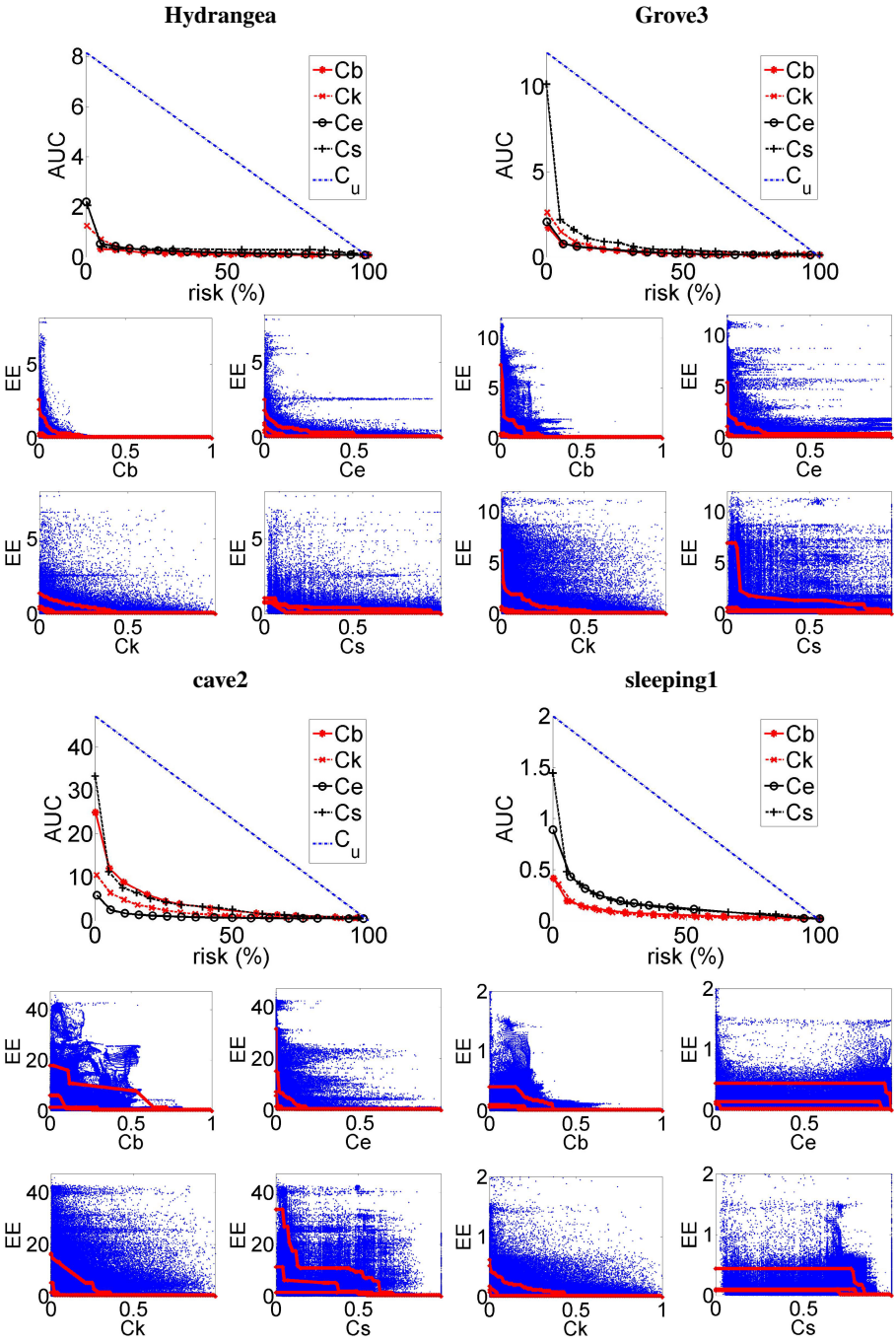


Fig. 5. RAUC and cm-EE scatter plots for Middlebury sequences Hydrangea and Grove3 and Sintel sequences cave2 and sleeping1

framework for assessing their quality. Being far from presenting a deep comparison between existing confidence measures, our examples clearly support the arguments about error sources given in this position paper. Our examples, also illustrate the usefulness of the concepts and tools introduced for facilitating research in confidence measure definition and evaluation of the best optical flow model. This encourages further research in comparison of confidence measure performance using a complete set of benchmark sequences for different applications.

Acknowledgements. Work supported by the Spanish projects TIN2009-13618, TIN2012-33116 and TRA2011-29454-C03-01 and first author by FPI-MICINN BES-2010-031102 program.

References

1. Baker, S., Scharstein, D., Lewis, J.P., et al.: A database and evaluation methodology for optical flow. IJCV 92(1), 1–31 (2011)
2. Barron, J.L., Fleet, D.J., Beauchemin, S.: Performance of optical flow techniques. IJCV 12(1), 43–77 (1994)
3. Bainbridge-Smith, R.G., Lane, A.: Measuring confidence in optical flow estimation. IET Electronics Letters 32(10), 882–884 (1996)
4. Kondermann, D., et al.: On performance analysis of optical flow algorithms. In: Dellaert, F., Frahm, J.-M., Pollefeys, M., Leal-Taixé, L., Rosenhahn, B. (eds.) Real-World Scene Analysis 2011. LNCS, vol. 7474, pp. 329–355. Springer, Heidelberg (2012)
5. Uras, S., Girosi, F., Verri, A., Torre, V.: A computational approach to motion perception. Biol. Cybern. 60, 79–97 (1988)
6. Singh, A.: An estimation-theoretic framework for image-flow computation. In: ICCV, pp. 168–177 (1990)
7. Kondermann, C., Kondermann, D., Garbe, C.S.: Postprocessing of optical flows via surface measures and motion inpainting. In: Rigoll, G. (ed.) DAGM 2008. LNCS, vol. 5096, pp. 355–364. Springer, Heidelberg (2008)
8. Simoncelli, E., Adelson, E., Heeger, D.: Probability distributions of optical flow. In: CVPR, pp. 310–315 (1991)
9. Lucas, B., Kanade, T.: An iterative image registration technique with an application to stereo vision. In: DARPA Image Understanding Workshop, pp. 121–130 (1981)
10. Bruhn, A., Weickert, J.: A confidence measure for variational optic flow methods. In: GPID, pp. 283–298 (2006)
11. Márquez-Valle, P., Gil, D., Hernández-Sabaté, A.: A complete confidence framework for optical flow. In: ECCV Workshops, pp. 124–133 (2012)
12. Kondermann, C., Mester, R., Garbe, C.S.: A statistical confidence measure for optical flows. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part III. LNCS, vol. 5304, pp. 290–301. Springer, Heidelberg (2008)
13. Kybic, J., Nieuwenhuis, C.: Bootstrap optical flow confidence and uncertainty measure. In: CVIU, pp. 1449–1462 (2011)
14. Mac Aodha, O., Humayun, A., Pollefeys, M., Brostow, G.J.: Learning a confidence measure for optical flow. IEEE PAMI (2012)
15. Evans, L.C.: Partial Differential Equations. American Mathematical Society (1998)
16. Horn, B., Schunck, B.: Determining optical flow. AI 17, 185–203 (1981)
17. Burden, R., Douglas Faires, J.: Numerical Analysis. Thompson (2005)
18. Shi, J., Tomasi, C.: Good features to track. In: CVPR, pp. 593–600 (1994)
19. Butler, D.J., Wulff, J., Stanley, G.B., Black, M.J.: A naturalistic open source movie for optical flow evaluation. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part VI. LNCS, vol. 7577, pp. 611–625. Springer, Heidelberg (2012)
20. Liu, C.: Beyond pixels: exploring new representations and applications for motion analysis. Ph.D. thesis (2009)