Evaluation of the Capabilities of Confidence Measures for Assessing Optical Flow Quality

Patricia Márquez-Valle Debora Gil Aura Hernàndez-Sabaté
Computer Vision Center
Edifici O, Campus UAB, Bellaterra, Spain
{pmarquez,debora,aura}@cvc.uab.cat

Abstract

Assessing Optical Flow (OF) quality is essential for its further use in reliable decision support systems. The absence of ground truth in such situations leads to the computation of OF Confidence Measures (CM) obtained from either input or output data. A fair comparison across the capabilities of the different CM for bounding OF error is required in order to choose the best OF-CM pair for discarding points where OF computation is not reliable. This paper presents a statistical probabilistic framework for assessing the quality of a given CM. Our quality measure is given in terms of the percentage of pixels whose OF error bound can not be determined by CM values. We also provide statistical tools for the computation of CM values that ensures a given accuracy of the flow field.

1. Introduction

Optical Flow (OF) is the input of a wide range of decision support systems such as car driver assistance, UAV guiding or medical diagnose. Discarding areas prone to have a large error in the computed flow is mandatory for ensuring reliable systems. In the absence of ground truth, the quality of the flow can only be obtained by a quantity computed from either sequences or the computed optical flow itself. These quantities are generally known as Confidence Measures, CM.

Confidence measures can be formulated from an analytic or a probabilistic point of view. Analytic approaches use the energy [3, 21] or the image structure (gradient magnitude [2], structure tensor [20]) as indicators of confidence. Energy-based approaches are linked to the capability of finding the energy minima and, thus, energy convexity. Meanwhile, 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 [10] or its variability with respect perturbations in the model [12]. 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 [15], and thus can be related to several sources of OF error.

Even if we have a proper confidence measure we still need a way to evaluate it. Given the large variety of OF methods and confidence measures, a fair comparison across them is not an easy task. 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 [1] compared seven different confidence measures for two image sequences. These results have been the 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 [11].

An early general attempt to define a type of confidence measures evaluation has been made by Bruhn et al. [3]. They validate the quality of confidence measures by means of 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 values. For each such threshold, the remaining average error is plotted. As explained in [16], 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 might result in possibly unfair comparisons between measures.

Aiming at a better comparison, the authors in [16] suggested a framework for confidence measure comparison based on its error bounding capabilities. The performance of confidence measures was assessed by computing the probability density function of having a decreasing dependency between flow errors and confidence measures. A main concern is the sampling of the 2D distribution space given by confidence and error values, which did not cover the whole space. In addition, it was not invariant to monotonic transformations of confidence measure values, which do not alter its error bounding capabilities. Attempting to solve the sampling problem, the same authors proposed in [17] scanning the 2D scatter given by confidence and error values by iteratively removing points having large values. Although the whole space was swept, the criteria for point removal was a critical point. Besides, invariance under transformations of the confidence measure was not achieved either.

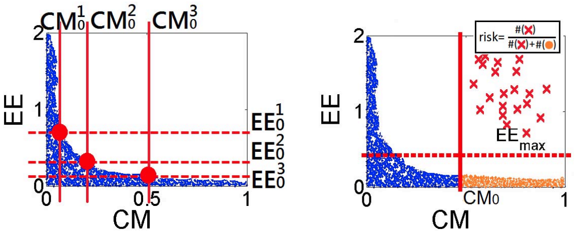
Based on the weak and strong points of existing comparison frameworks, this paper presents an statistical probabilistic framework for assessing the quality of a given CM for bounding the error of a particular OF method. Our measure is given in terms of the percentage of pixels (called risk) which bound can not be determined by CM values. The profile of the plots given by the risk over CM percentiles provide information about the capabilities of each pairing OF-CM for predicting the percentage of pixels with unbounded error. In order to account for monotonically increasing CM transformations, CM-error scatter plots are sparsified as in [3] using CM percentiles. We call these plots Sparse-Density Plots, SDP.

Confidence intervals of the risk computed over frames belonging to benchmark databases, can be used to discard bad areas in sequences presenting similar visual features. In this paper we have chosen three different OF methods and four representative confidence measures to validate our quality framework on the Sintel database [5]. Results indicate that the energy-based measure [3] is well-suited for discarding OF erroneous outputs, providing that assumptions are met.

2. Evaluation Framework

In an ideal case, we would expect the values of a confidence measure to be correlated to the flow End-point Error, EE. 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. 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 [6].

In order that a measure is useful for bounding errors, the scatter plot between the measure and end-point errors should show a monotonic tendency. In other words, if CM values are small, then, OF error is not bounded and it can take any value. Meanwhile, for large CM values, OF error should be bounded so that the output data is reliable. Figure 1 (a) shows a representative scatter plot for CM-EE values and illustrates the expected decreasing behavior between CM and error.



(a) (b)

Figure 1. (a) Scatter plot between CM and EE showing a good profile for the confidence measure. Red vertical lines show different CM_0^i, and red dashed lines the respective EE_0^i bound. (b) Concept of error bound and risk. Red crosses correspond to points that a confidence measure can not bound. For each percentile (in this case vertical line at CM_0) risk is the percentage of points above the EE_{max}.

Taking into account the expected relation between CM and error values, the evaluation of CM quality should assess its capabilities for bounding OF error. This can be achieved by exploring the decreasing profile of CM-error scatter plots.

2.1. Sparse Density Plots

In the best case, 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. In probabilistic terms, this implies that the following conditional probability is zero:

$$\forall CM_0, \exists EE_0 \text{ s.t. } P(EE > EE_0 | CM > CM_0) = 0 \quad (1)$$

Figure 1 (a) illustrates EE error bounds for different CM_0 values in the case of a perfect relationship between CM and EE. Vertical lines correspond to several CM_0 values and horizontal lines the best EE_0 bound for such CM_0 values.

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

$$\exists CM_0 \text{ s.t. } \forall EE_0, P(EE > EE_0 | CM > CM_0) > 0 \quad (2)$$

We define the risk of a confidence measure as the proportion of points, ρ, which bound can not be determined by CM values. We note that ρ is, in fact, a function of the threshold values CM_0. Since decision support systems usually require a minimum accuracy, we compute ρ in terms of the maximum allowed error:

$$\rho(CM_0) := P(EE > EE_{max} | CM > CM_0) \quad (3)$$

for EE_{max} the maximum error allowed. It should be clear that the lower the risk, the higher the power for bound prediction CM has. The scatter plot in fig. 1(b) showing

CM versus EE illustrates the concept of risk. The vertical red line represents the threshold for CM and the horizontal dashed red line the bound on EE given by EE_{max}. For each CM_0 its risk is given by the percentage of points on the upper right square defined by the two lines, which correspond to the red crosses on the upper part of the plot.

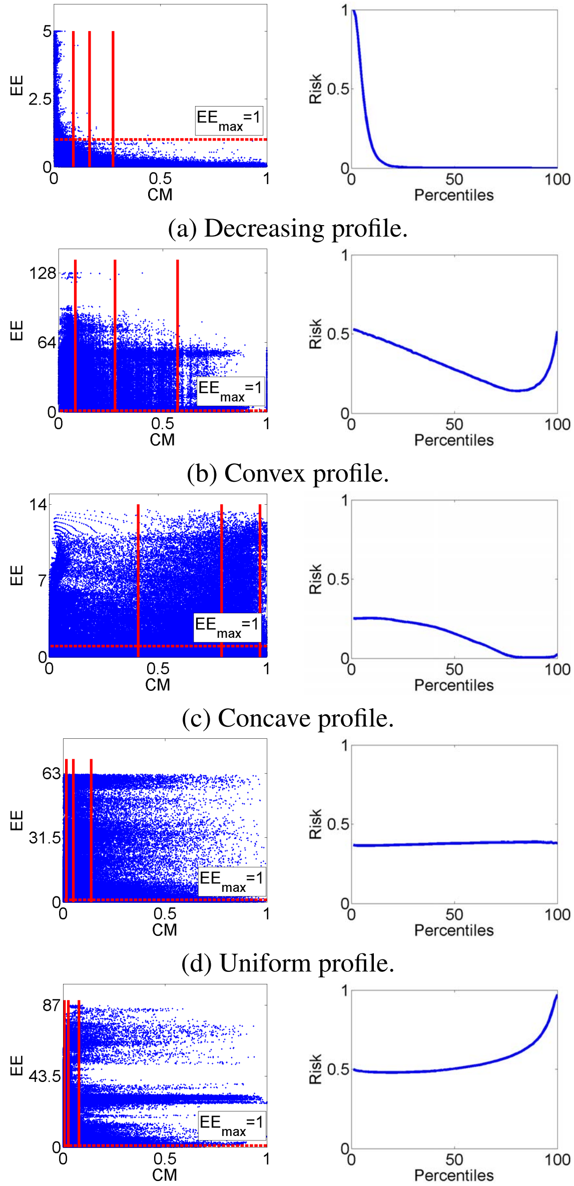
Under the considerations above it should be clear that the profile of the plots given by the risk as a function of CM, $(CM, \rho(CM))$ provide information about CM error bound capabilities. It is worth noticing that CM error bounding capabilities are invariant under monotonically increasing transformations of CM, which only modify the value CM_0 achieving a given risk. It follows that CM scatter plots should be sampled so that the plots $(CM_0, \rho(CM_0))$ are invariant under monotonically increasing transformations of CM. This can be achieved by using the percentiles of CM distribution, namely $prct_{CM}$, instead of CM values. We define our Sparse-Density Plots (SDP) as the plots given by:

$$(prct_{CM}, \rho(prct_{CM})) \tag{4}$$

Figure 2 shows the main SDP profiles ranged from best to worst capabilities for error bounding. Left column corresponds to the scatter plot CM-EE with the percentiles $\{0.25, 0.5, 0.75\}$ marked in red lines, and the EE_{max} in dashed red line at 1, while right column shows the corresponding SDP. A confidence measure is able to completely bound OF error if SDP has an strictly decreasing profile and reaches the zero value for some $prct_{CM}$, like the profile shown in fig2 (a). In such case, pixels belonging to the upper percentile $[prct_{CM}, 1]$ have no risk at all, so its error is bounded. Plots shown in figs.2 (b) and (c) come from the most usual CM behaviors. In the first case (fig.2 (b)), there is a small quantity of points where the error is never bounded by CM values. This introduces an increasing profile at the end of SDP graphics. In the second case (fig.2 (c)), there is a group of pixels with unbounded errors in the first CM percentiles but for higher percentiles, the error is completely under control. Finally, figs.2 (d) and (e) show the worse cases, in the sense that CM is not related to OF error. The constant profile of fig.2 (d) indicates that the $CM - EE$ distribution is uniform and, thus, EE can take any value regardless of CM. The case shown in fig.2 (e) is even worse. It has a behavior opposite to the expected one as large CM values have an unbounded error.

2.2. Risk Prediction

The procedure described so far can only be computed for a representative sample extracted from sequences with ground truth. For the generalization to any sequence, statistical inference should be applied. In this framework, we should determine a confidence interval for the risk given a percentage of points we would like to keep. In order to do so, the variability of SDP across a sample of representative sequences should be as low as possible [19]. In this context, the most relevant feature of confidence measures is not only a decreasing SDP pattern but also a stable behavior across different sequences.



(a) Decreasing profile.

(b) Convex profile.

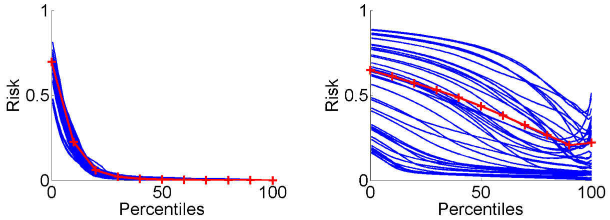
(c) Concave profile.

(d) Uniform profile.

(e) Increasing profile.

Figure 2. Representative examples of different SDP, ranged from best to worst capabilities for error bounding. Left column shows to the scatter plots (CM vs EE). Vertical red lines correspond to the percentiles $0.25, 0.5, 0.75$ and horizontal red line indicates the $EE_{max} = 1$. Right column shows the respective SDP.

For each $prct_{CM}$, the risk values $\rho(prct_{CM})$ taken across a sample of frames presenting comparable appearance and dynamic conditions define a random variable X_{prct}. The one-sided confidence interval [7] for X_{prct} mean gives a punctual upper bound for the risk at each given



(a) (b)

Figure 3. Risk prediction capabilities for sets presenting low, (a), and large, (b), variability in SDP profiles.

percentile $prct_{CM}$. If $\mu(X_{prct})$, $\sigma(X_{prct})$ are, respectively, the sample mean and variance for X_{prct} computed on a sampling of size N ($N > 30$), then the interval at confidence level $1 - \alpha$ is given by:

$$[0, \mu(X_{prct}) + t_\alpha \sigma(X_{prct})] = [0, \rho_{X_{prct}}] \tag{5}$$

for t_α the value of a T-Student distribution with N-2 degrees of freedom having a cumulative probability equal to $1 - \alpha$ [7, 18].

If we consider the upper bounds, $\rho_{X_{prct}}$, as a function of CM percentiles $prct_{CM}$, we get a curve:

$$\Gamma := \Gamma(prct_{CM}) = (prct_{CM}, \rho_{X_{prct}})$$

that should provide an upper bound for the error risk of new incoming sequences with similar conditions as the sample used to compute (5). That is, once a $prct_{CM}$ of pixels have been removed, the error of the remaining ones should be under EE_{max} with probability $\rho_{X_{prct}}$.

The curve Γ actually provides a proper bound if SDP profiles present a moderate variability across sequence frames. Otherwise, the risk can not be bounded by means of Γ values. Figure 3 illustrates the computation of Γ for a set presenting small variability (fig. 3(a)) and a set with a large variability (fig. 3(b)). The curve Γ is shown in red and the testing samples in blue. We observe that in the case of small variability the distribution of the blue curves concentrates around Γ, while it has a larger dispersion for the less predictable set.

It follows that the capability of a pair OF-CM for risk bounding can be assessed using the dispersion of a set of test SDP curves around the curve Γ. The dispersion can be expressed using the difference function:

$$\begin{aligned} Z_{Diff}(prct_{CM}) &:= \Gamma(prct_{CM}) - SDP(prct_{CM}) \\ &= \rho_{X_{prct}} - \rho(prct_{CM}) \end{aligned} \tag{6}$$

For each $prct_{CM}$, the function $Z_{Diff}(prct_{CM})$ is a continuous function that takes values in the range $[-1, 1]$. Negative values indicate that SDP curves are actually bounded by Γ, while positive ones represent the risk of deviation from the bound given by Γ values. The function $Z_{Diff}(prct_{CM})$ taken across a sampling set of sequences defines a continuous variable. In this context, we consider that Γ is a proper bound if the average of $Z_{Diff}(prct_{CM})$, namely $\mu_{Z_{Diff}}$, is close to zero or negative. This can be statistically checked using the following one-tailed t-test [18]:

$$\begin{aligned} H_0 : &\quad \mu_{Z_{Diff}} \geq \mu_0 \\ H_1 : &\quad \mu_{Z_{Diff}} < \mu_0 \end{aligned} \tag{7}$$

If the null hypothesis H_0 is rejected, then the upper bound Γ is able to predict the error risk for sequences with the same features up to an average deviation equal to μ_0. The value of μ_0 represents the percentage of risk increase we admit in our predictions. Thus, it varies according to the further use of flow computation and, in particular, it is set by the robustness of the decision support system to outliers.

3. Experimental Settings

The goal of our experiments is to show the applicability of the presented framework for selecting OF-CM pairs able to predict the risk for a given type of sequences. In order to do so, two experiments have been carried out:

1. **Explorative analysis of predictable sequences.** The validity of the bound given by (5) is assessed at level sequence. However, not all sequences are proper to predict a bound because they are either too good (thus the resulting flow field has no errors) or too bad (they do not fulfil theoretical model assumptions). For that reason, the first step is to carry on an explorative analysis of predictable sequences before assessing error bound at sequence level. The predictability of sequences is assessed by matching the profile of the Γ curves to the main SDP profiles shown in fig.2.

2. **Assessment of error bound at sequence level.** For each OF method, unpredictable sequences have been removed for assessment of the error bound capabilities. For each sequence, Γ curves have been computed using a sampling of 20 random frames over a uniform sampling of $prct_{CM}$ given by $\{0, 0.1, \ldots, 1\}$. The variable Z_{Diff} is defined by considering concatenating values for all sequences. Tests are done at a significance level $\alpha = 0.05$ [18]. The risk increase has been set to $\mu_0 = 0.05$, which implies that we assume up to a 5% of deviation from the risk predicted by Γ.

In order to cover as much methods and sequence features as possible, we have chosen the Sintel database [5], three representative OF methods, and four confidence measures with different grounds.

Database. We have selected 17 sequences from the Sintel Database [5] with at least 40 frames in order to perform the experiments. The Sintel database contains sequences with large motion, specular reflection, motion blur, defocus blur and atmospheric effects and, thus, covers a complete bunch of sequence features.

Optical flow algorithms. In order to asses the range of applicability of our framework, we have applied it to the following representative and state of the art optical flow methods [1]:

- Combined local-global method (**CLG**) [4]: This method combines Lucas-Kanade data term [14] with an L^2 norm smoothness term. The method CLG has been chosen due to its simplicity on the formulation and also because this method has as data term a solvable equation, that is, the data term can solve the optical flow on its own.

- Horn-Schunck method (**HS**) [8]: This is the classic approach that uses OF brightness constancy equation with an L^2 norm smoothness term. The method HS has been chosen because it is the original formulation for variational OF techniques. In addition, due to its simplicity on the formulation, we can control better the different error sources and thus it facilitates a further analysis of confidence measures performance.

- Classic-NL method (**NL**) [22]: This total variation method uses the L^1 norm to combine OF brightness constancy assumption with the smoothness term. The method NL has been chosen because it is one new implementation that uses the L^1 norm to avoid over-regularization of the OF computation.

Confidence Measures. In order to find optimal confidence measures for each OF method, we have considered four CM with different grounds:

- Image structure (C_k). From all measures based on image structure [2], we selected the condition number of the data-term system defined in [16].

- Energy (C_e). The confidence measure is computed by evaluating the flow field over the functional as described in [3].

- Statistical (C_s). It assesses the computed optical flow calculating the local variability by means of the Mahalanobis distance between the computed vector and the distribution given by the surrounding ones [9].

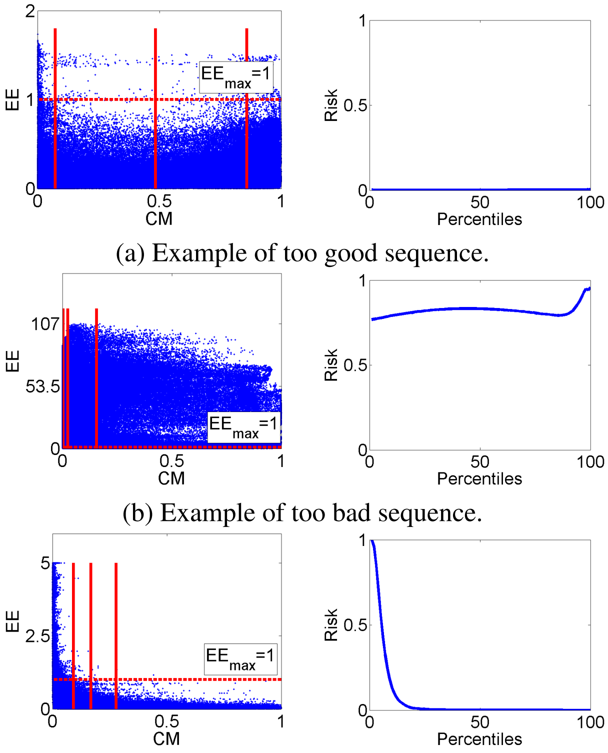
- Bootstrap (C_b). It measures OF variability with respect to perturbations in the model [12].

Therefore, we have $3 \times 4 = 12$ possible OF-CM pairs.

4. Results and discussion

4.1. Explorative analysis of predictable sequences

A first analysis of the predictability of the sequences is summarized in table 1. For each sequence and each pair OF-CM, a label ranging from -1 to 3 is assigned to the profiles of the trained upper bound. The labels are assigned following the opposite order of the plots shown in figure 2 from the worst profile (-1) to the best one (3). As well, those sequences that have a good profile and the upper bound is below 1 for all pixels have been labeled by a 3*. To make the table more readable, we have assigned a different color to each label.



(a) Example of too good sequence.

(b) Example of too bad sequence.

(c) Example of predictable sequence.

Figure 4. SDP profiles for different kind of sequences: prediction not necessary (a), non-predictable (b) and predictable (c). Left column shows to the scatter plots (CM vs EE). Vertical red lines correspond to the percentiles $0.25, 0.5, 0.75$ and horizontal red line indicates the $EE_{max} = 1$. Right column shows the respective SDP.

Focusing on columns, we can observe that there are three kind of sequences: *too good sequences* (labeled by 3*), *too bad sequences* (labeled by -1) and *predictable* ones.

Too good sequences: Current optical flow methods are able to accurately solve the flow field of sequences fulfilling the method theoretical requirements (brightness constancy, small displacements, etc). There are some sequences that met such requirements and thus they had not only a good profile for all pairs OF-CM but also a very low upper bound

[1] Using the free source code from [13] for the CLG method and [22] for HS and NL methods.

		alley_1	alley_2	ambush_5	ambush_7	bamboo_1	bamboo_2	bandage_1	bandage_2	cave_2	cave_4	market_2	market_5	mountain_1	shaman_2	shaman_3	sleeping_1	sleeping_2	Label average
CLG	ck	0	3*	-1	0	0	0	0	0	-1	-1	0	-1	0	0	-1	3*	3*	0
	cb	3	3*	-1	3	3	3	3	3	-1	-1	3	-1	2	3	-1	3*	3*	2,9
	ce	3	3*	-1	3	3	3	3	3	-1	-1	2	-1	2	2	-1	3*	3*	2,7
	cs	1	3*	-1	3	2	3	3	2	-1	-1	3	-1	2	2	-1	3*	3*	2,3
HS	ck	0	3*	0	0	0	0	0	0	-1	0	0	-1	0	0	0	3*	3*	0
	cb	1	3*	1	2	3	3	3	3	-1	3	1	-1	3	3	3	3*	3*	2,4
	ce	3	3*	3	3	3	3	3	3	-1	3	3	-1	1	3	3	3*	3*	2,8
	cs	1	3*	-1	3	2	3	3	2	-1	2	3	-1	0	1	2	3*	3*	1,8
NL	ck	0	3*	0	0	0	0	0	0	-1	0	0	-1	-1	0	-1	3*	3*	0
	cb	3	3*	-1	-1	3	3	-1	-1	-1	-1	3	-1	-1	3	-1	3*	3*	1
	ce	3	3*	2	3	3	3	3	3	-1	3	2	-1	-1	3	-1	3*	3*	2,8
	cs	1	3*	2	1	2	2	3	1	-1	1	3	-1	-1	3	-1	3*	3*	1,9

Table 1. Profile labels for each pair OF-CM.

of the EE ($alley_2$, $sleeping_1$, and $sleeping_2$). Since the error of the optical flow for those sequences is below 1 pixel for almost all pixels, the risk is almost 0 for all percentiles. Thus, the SDP does not provide additional information, and further prediction is not necessary. Therefore, these sequences are removed from further analysis. Figure 4(a) shows an illustrative example of this kind of sequence. On the left column, the scatter plot shows that most of points are below EE_{max}. This means that the error is subpixel and thus, for each percentile, the risk is almost 0, as we can observe in the plot on the right hand side of the figure.

Too bad sequences: Different OF methods require specific assumptions in order to properly perform the flow field. In case sequences do not fulfil such requirements, errors are arbitrarily large. In this case, none of the CMs is able to relate to the error and these sequences have to be removed from further analysis. This is the case of the sequences which scored -1 for all CM and an OF method, shown in figure 1. For instance, CLG is based on Lucas-Kanade, thus, its performance drops in case images do not have enough texture or corners, like $ambush_5$, $cave_4$ or $shaman_3$. In the case of NL, the use of an approximation to the L^1 norm (which can not be derived near zero) disturbs results in case images have large areas of uniform intensity, like $mountain_1$ and $shaman_3$. Besides, fast motion introduces sudden changes in appearance and new objects and occlusions abruptly appear into the scene ($market_5$) or blurs too much the image ($cave_2$), making any OF method fail. As well, in the case of $market_5$, illumination changes violate brightness constancy constrain. Whether optical flow assumptions are met should be checked a priori using image processing. Figure 4(b) shows an illustrative example of this kind of sequence. The scatter plot on the left hand side of the figure, we can observe that most of the points are above EE_max, and thus the risk is high (shown on the right hand side of the figure). As well, and most important, the density of the scatter plot is accumulates on the upper percentiles (marked in vertical red lines), resulting an increasing risk profile, which is not able to be predicted.

Predictable sequences: The remaining sequences obtain different scores along the different pairs OF-CM, and thus, there exists at least one pair OF-CM that can predict the risk. This set of sequences are the candidates to carry on the assessment of error bound at sequence level. Figure 4(c) shows an illustrative example of this kind of sequence. The scatter plot on the left hand side shows the density of points is accumulated on the lower percentiles (marked in red vertical lines), and most of them are below 0.25 percentile. This results in a decreasing profile of the curve, shown on the right hand side of the figure, and thus, the risk can be predicted.

Once we have removed non-predictable sequences and sequences where prediction is not necessary, we can observe that different measures have different performances according to methods or sequences. Given that the average of our score indicates its prediction capabilities, pairs below 2 will be discarded. In this context, measure C_k scores 0 for all methods because it is too restrictive and discards all pixels for this database. The confidence measure C_b is successful when the data-term of the flow algorithm can resolve optical flow by itself (without the regularity term), this holds for L^2 approaches (and specially for CLG scheme) but not for total variation methods such as NL. In this case, its average score is 0.87 and, thus, it should be discarded. The measure C_e is adequate if model assumptions are met, thus, it is the best performer for our selected data-set because non-predictable frames coincide with frames failing to met the optical flow algorithm requirements. Finally, C_s depends more on the nature of optical flow and achieves better results in the presence of patches presenting regular motion. It follows that its average drops below 2 for HS and NL, which are the OF methods that include more variability in sequences.

For that reasons, we consider as a proper candidates the following pairs OF-CM: CLG-C_b, CLG-C_e, HS-C_b, HS-C_e, NL-C_e. The tests explained in subsection 2.2 will serve to check the predictability of the error of those pairs for sequences with similar features.

4.2. Assessment of error bound at sequence level

Table 2 summarizes the Z_{Diff} t-test statistics associated to each OF-CM pair. We report the null hypothesis (with 1 if it is rejected), the p-value and the confidence interval upper bound of the difference test.

For all cases, risk could be predicted with a deviation below 10% (as CI upper bound in last column are below .08). However, none of the measures succeeded in predicting HS risk with a deviation less than 5%. This is mainly attributed to a large variability in HS performance, which is very sensitive to the theoretical assumptions (such as sudden changes in OF spatial distribution or smooth image appearance). This introduces large variation in SDP plots as illustrated in fig. 5 which shows SDP and Γ for the $ambush_7$ sequence. This is a representative case of a sequence without all frames meeting the assumptions required for the OF method. Like in real world conditions, this is a common issue along most Sintel sequences and suggests a different basic sampling (image patches, for instance) other than sequence frames for computing statistics. The measure C_s also failed to achieved the desired performance, probably due to its high dependance on motion patch regularity.

The only pairs that are good candidates to predict error risk with a deviation under 5% are CLG-C_b, CLG-C_e and NL-C_e. This is a sensible result given that the set of test sequences that were considered for this experiments fulfilled the theoretical assumption required for good performance of the OF methods (see the discussion carried out at the end of Section 4.1).

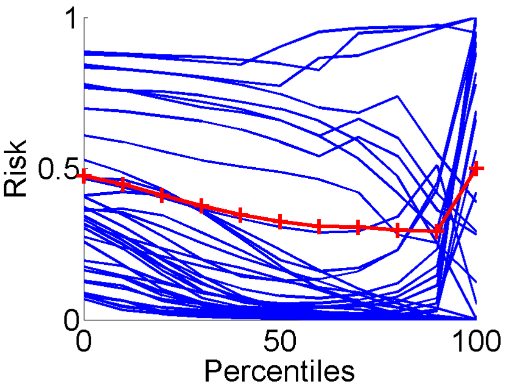


Figure 5. SDP profiles and curve Γ for the sequence $ambush_7$.

		H	p-val	CI
CLG	Cb	1	4.9977e-05	0.0460
CLG	Ce	1	0.0073	0.0485
CLG	Cs	0	0.9997	0.0631
HS	Cb	0	1	0.0798
HS	Ce	0	1.0000	0.0601
NL	Ce	1	0.0022	0.0481

Table 2. Difference test.

5. Conclusions and Further Research

This paper presents statistical and probabilistic tools for validating the capabilities of confidence measures for assessing OF quality in the absence of ground truth. Confidence measures are evaluated in the measure that they establish a threshold ensuring a bound on OF error up to a percentage of pixels (SDP plots). We also describe the statistics needed to compute the threshold (the Γ curve) of the confidence measure that ensures a given accuracy of the flow field. Our framework has been validated on the Sintel database [5], by three representative OF methods, and four confidence measures with different grounds.

Our experiments indicate that measures based on either local image structure [2] or local motion regularity [9] are not the best suited for predicting OF error risk, at least for the considered OF methods. Energy-based [3] and bootstrap [12] measures are better candidates, as far as, sequences match some assumptions. In particular, the bootstrap is suitable for CLG methods, while the energy-based could predict error risk for a wider range of variational methods. None of the confidence measures could predict HS risk, probably due to a bad definition of the samples used to compute Γ.

This preliminary work constitutes a first new effort in the use of statistical and probabilistic tools for the evaluation of the capabilities of CM for predicting OF error in decision support systems. However, more research is needed in order to fully validate our framework as a solid methodology for the implementation of OF error prediction strategies.

The performance of our methodology depends on several factors. On the one hand, it depends on the variability of the dynamical appearance of the frames taken in, both, the training set used to compute Γ and new incoming sequences. On the other hand, the error of a pair OF-CM can be predicted provided that sequences fulfill some assumptions. In case this assumption are not met, SDP profiles have a significant large variability that can not be properly modeled by the Γ curves. Therefore the risk of this kind of sequences can not be predicted and should be excluded a priori.

In our experiments, we visually identified the theoretical assumptions. This already validates our methodology for some application fields having very controlled acquisition conditions, such as medical imaging. However, in order that our system can effectively run on an arbitrary decision support application (such as ADAS), whether OF assumptions are met should be checked a priori using image processing.

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] R. Bainbridge-Smith, A. Lane. Measuring confidence in optical flow estimation. *IET Electronics Letters*, 32(10):882–884, 1996. 1

[2] J. L. Barron, D. J. Fleet, and S. S. Beauchemin. Performance of optical flow techniques. *IJCV*, 12(1):43–77, 1994. 1, 5, 7

[3] A. Bruhn and J. Weickert. A confidence measure for variational optic flow methods. In *Geometric Properties for Incomplete Data*, pages 283–298, 2006. 1, 2, 5, 7

[4] A. Bruhn, J. Weickert, and C. Schnörr. Lucas/Kanade meets Horn/Schunck: Combining local and global optic flow methods. *IJCV*, 61(2):221–231, 2005. 5

[5] D. Butler, J. Wulff, G. Stanley, and M. Black. A naturalistic open source movie for optical flow evaluation. In *ECCV*, 7577, pages 611–625. Springer-Verlag, 2012. 2, 4, 7

[6] W. Cheney and D. Kincaid. *Numerical Mathematics and Computing, Sixth edition*. Bob Pirtle, USA, 2008. 2

[7] R. Fisher. *Statistical Methods and Scientific Inference*. Oliver and Boyd, 1956. 3, 4

[8] B. Horn and B. Schunck. Determining optical flow. *AI*, 17:185–203, 1981. 5

[9] C. Kondermann, D. Kondermann, B. Jähne, and C. S. Garbe. An adaptive confidence measure for optical flows based on linear subspace projections. In *DAGM-Symposium*, volume 4713 of *Lecture Notes in Computer Science*, pages 132–141. Springer, 2007. 5, 7

[10] C. Kondermann, R. Mester, and C. Garbe. A statistical confidence measure for optical flows. In *ECCV*, pages 290–301, 2008. 1

[11] D. Kondermann, S. Abraham, G. Brostow, W. Förstner, S. Gehrig, A. Imiya, B. Jähne, F. Klose, M. Magnor, H. Mayer, et al. On performance analysis of optical flow algorithms. *Outdoor and Large-Scale Real-World Scene Analysis*, pages 329–355, 2012. 1

[12] J. Kybic and C. Nieuwenhuis. Bootstrap optical flow confidence and uncertainty measure. *Computer Vision and Image Understanding*, pages 1449–1462, 2011. 1, 5, 7

[13] C. Liu. *Beyond pixels: exploring new representations and applications for motion analysis*. PhD thesis, Cambridge, MA, USA, 2009. 5

[14] B. Lucas and T. Kanade. An iterative image registration technique with an application to stereovision. In *DARPA IU Workshop*, pages 121–130, 1981. 5

[15] O. Mac Aodha, A. Humayun, M. Pollefeys, and G. J. Brostow. Learning a confidence measure for optical flow. *IEEE PAMI (early access articles)*, PP, 2012. 1

[16] P. Márquez-Valle, D. Gil, and A. Hernàndez-Sabaté. A complete confidence framework for optical flow. In *ECCV Workshops*, volume 7584 of *LNCS*, pages 124–133. Springer, 2012. 1, 5

[17] P. Márquez-Valle, D. Gil, A. Hernàndez-Sabaté, and D. Kondermann. When is a confidence measure good enough? In *ICVS*, pages 344–353. Springer Link, 2013. 2

[18] D. Moore. *The basic practice of statistics (6th edition)*. 2006. 4

[19] P. Newbold, W. Carlson, and B. Thorne. *Statistics for Business and Economics*. Pearson Education, 2007. 3

[20] J. Shi and C. Tomasi. Good features to track. pages 593–600, 1994. 1

[21] A. Singh. An estimation-theoretic framework for discontinuous flow fields. In *ICCV*, pages 168–177, 1990. 1

[22] D. Sun, S. Roth, and M. J. Black. Secrets of optical flow estimation and their principles. In *CVPR*, pages 2432–2439, 2010. 5