A Confidence Framework for the Assessment of Optical Flow Performance

A dissertation submitted by **Patricia Márquez Valle** at Universitat Autònoma de Barcelona to fulfil the degree of **Doctora en Informàtica**.

Bellaterra, Juliol 2015

Advisors: **Dra. Debora Gil Resina i Dra. Aura Hernàndez i Sabaté**
Dep. Ciències de la Computació i Centre de Visió per Computador
Universitat Autònoma de Barcelona

Thesis Committee
Dra. Laura Igual Muñoz
Dep. Matemàtica Aplicada i Anàlisis
Universitat de Barcelona
Dra. Maria Mercè Farré Cervelló
Dept. Matemàtiques
Universitat Autònoma de Barcelona
Dr. Naveen Onkarappa
Samsung R&D Center
Dr. Jordi Gonzalez Sabaté
Dept. Ciències de la Computació
Universitat Autònoma de Barcelona
Dr. Christoph S. Garbe
Interdisciplinary Center for Scientific Computing (IWR)
University of Heidelberg

This document was typeset by the author using LaTeX 2_ε.

The research described in this book was carried out at the Computer Vision Center, Universitat Autònoma de Barcelona.

ISBN 978-84-943427-2-1

A tu.

Acknowledgments

Un dia vas a fer una inscripció per a un màster i no saps com però acabes amb una beca per a fer un doctorat i un projecte que esperes tindrà un final no massa llunyà. Els anys passen i no acabes de saber com finalitzar això que un dia vas decidir començar... (bé, ho marca la beca que té final definit, 4 anys!) Sort he tingut de tindre dues tutores que m'han espavilat per a què enllesteixi amb la tesi, si no encara estaria avorrint-se en un ordinador! Així doncs, especial agraïment a les meves tutores, la Debora i l'Aura, per la paciència i dedicació durant aquests 4 anys que s'han acabat convertint en gairebé 5, qui ho diria! Sense vosaltres aquesta tesi no s'hauria acabat mai!

Debora, amb tu he après que la festa és una cosa innata a la persona i que no importa l'edat ni la professió. Ara que sembla que no ens haurem d'aguantar en més congressos, espero que mantinguis el llistó ben alt de *white nights*!

Aura, amb tu he après que encara que sembli impossible reunir-se amb algú sempre hi ha un forat per a trobar-se. Després d'aquests dos últims mesos de feina condensada, es faran estranys els dilluns sense el Colacao i menjar a domicili!

Durant aquests anys no només m'han aguantat les meves tutores, així doncs anem a fer memòria....

Marc, recordo el dia de la inscripció del màster, vam anar amb les idees clares i un gran NO VULL FER UN DOCTORAT escrit al front, de cop i volta aquell home ens va dir que hi havia beques per entrar a fer un doctorat al CVC i ens va omplir el cap de pardals. Aquell mateix dia estàvem enviant mails als caps de grup a veure si algú ens volia per a fer un doctorat. I vam tindre la desgràcia i la sort de què hi va haver gent que ens va voler. La desgràcia, suposo, ha estat haver de fer el doctorat, dic suposo perquè com tot en aquesta vida, ha tingut coses bones i no tan bones (de no tan bones *"el final"* que s'ha convertit en un túnel fosc i sense sortida). La sort va ser que ens van agafar als dos i que t'he pogut aguantar durant aquests anys, així que prepara't que toca festa grossa per celebrar que acabem el doctorat!

El grup on vaig anar a parar per a fer la tesi semblava que no tingués gaire gent, però al final si m'hi paro a pensar, déu n'hi do! Gràcies Albert! Amb tu vaig aprendre que les actualitzacions dels ordinadors no m'agraden, tot i així em fascina la capacitat que tens per a calibrar la imatge de les pantalles de l'ordinador. Va ser pesat etiquetar les imatges del cor, però va ser l'única part d'imatge mèdica que vaig tocar, així que en el fons va estar bé. Gràcies Carles per llevar-te a temps per anar a veure la presentació del congrés, sé que va ser dur! I gràcies per resoldre dubtes

d'última hora i per ajudar sempre que ha fet falta! Gràcies Antonio per ajudar amb les imatges de broncoscòpia, la tesi queda millor ara, ha estat una col·laboració curta però intensa, no em maleeixis quan hagis de fer anar el meu codi! Gràcies Francesco per ser part del grup i compartir penes amb les ANOVA's! A veure si ens tornem a trobar en una piscina o al mar! I gràcies a tota la resta de gent del grup per estar per allà, sempre s'agraeix que hi hagi algú!

Aquests anys de tesi els he passat en un soterrani, sembla mentida, però un li acaba agafant *carinyo*. Gràcies a tots els del soterrani per ser allà a baix! Especial gràcies a l'Anjan, la Núria i la Yainuvis per compartir aquest racó amb llum artificial on no ha funcionat mai bé la calefacció.

Thanks to all the people I met during my research stays in Germany and in the Netherlands. I had a great time in both countries, thanks to all friends I made. Thanks Rudolf for giving me the chance of visiting different research centers. Thanks Daniel for sharing your knowledge about caffeine effects and side effects. Thanks Christian for sharing Kas Limon and Cacaolat. Thanks Hans and Luc, I enjoyed working with you. Thanks Hanne for your warm welcome. Thanks Lena for the Poker game, I had never won that game before!

I arriba el torn del tribunal! Thanks Naveen for the Optical Flow evenings! It was great to find someone in CVC that was working on that. Gràcies Mercè per ensenyar-me estadística a la carrera, amb el que em va costar aprovar l'assignatura, qui diria que acabaria fent una tesi on, a més a més de tests d'hipòtesis, hi apareixen ANOVA's! Gràcies Laura per formar part del tribunal. Sempre ens anem trobant per aquí i per allà, ara fa temps que no ens veiem, així que suposo que ja toca, aquest cop sí, amb data i hora! So, thanks to the thesis committee for reading the thesis and see you on the 27th of July!

I ara sí, per acabar, gràcies a totes aquelles persones que s'han creuat en la meva vida, sense cap mena de dubte, en més o menys mesura han contribuït en aquesta tesi, ja que sense ells no seria qui sóc ni estaria on estic ara.

P.D.1.: Debora, Aura, és la primera vegada que em poso a escriure i escric tant! M'he hagut de controlar i tot per no expandir massa, llàstima que només hagi sortit la vena escriptora en els acknowledgements!

P.D.2.: I que no m'oblidi d'agrair qui ha invertit en mi per a fer aquesta tesi:

This work was supported by FPI-MICINN BES-2010-031102 program from the Ministry of Economy and Competivity of Spain.

Abstract

Optical Flow (OF) is the input of a wide range of decision support systems such as car driver assistance, UAV guiding or medical diagnose. In these real situations, the absence of ground truth forces to assess OF quality using quantities computed from either sequences or the computed optical flow itself. These quantities are generally known as Confidence Measures, CM. Even if we have a proper confidence measure we still need a way to evaluate its ability to discard pixels with an OF prone to have a large error. Current approaches only provide a descriptive evaluation of the CM performance but such approaches are not capable to fairly compare different confidence measures and optical flow algorithms. Thus, it is of prime importance to define a framework and a general road map for the evaluation of optical flow performance.

This thesis provides a framework able to decide which pairs *"optical flow - confidence measure"* (OF-CM) are best suited for optical flow error bounding given a confidence level determined by a decision support system. To design this framework we cover the following points:

- **Descriptive scores.** As a first step, we summarize and analyze the sources of inaccuracies in the output of optical flow algorithms. Second, we present several descriptive plots that visually assess CM capabilities for OF error bounding. In addition to the descriptive plots, given a plot representing OF-CM capabilities to bound the error, we provide a numeric score that categorizes the plot according to its decreasing profile, that is, a score assessing CM performance.

- **Statistical framework.** We provide a comparison framework that assesses the best suited OF-CM pair for error bounding that uses a two stage cascade process. First of all we assess the predictive value of the confidence measures by means of a descriptive plot. Then, for a sample of descriptive plots computed over training frames, we obtain a generic curve that will be used for sequences with no ground truth. As a second step, we evaluate the obtained general curve and its capabilities to really reflect the predictive value of a confidence measure using the variability across train frames by means of ANOVA.

The presented framework has shown its potential in the application on clinical decision support systems. In particular, we have analyzed the impact of the different image artifacts such as noise and decay to the output of optical flow in a cardiac diagnose system and we have improved the navigation inside the bronchial tree on bronchoscopy.

Abstract (català)

L'*Optical Flow* (OF) és l'*input* d'una gran varietat de Sistemes de Suport a Decisions (DSS) com ara assistència a la conducció, guia UAV o diagnosis mèdic. En aquestes situacions, l'absència de *ground truth* ens obliga a avaluar la qualitat de l'OF calculat mitjançant quantitats calculades a partir de les seqüències o bé a partir del mateix OF. Aquestes quantitats es coneixen generalment com a Mesures de Confiança (CM). Encara que tinguem una mesura de confiança, necessitem alguna eina per tal d'avaluar la seva capacitat per descartar píxels de la imatge que tenen tendència a tindre error. Els mètodes actuals només aporten una avaluació descriptiva del rendiment de les CM, el problema és que aquests mètodes no són capaços de comparar equitativament les diferents CM i OF. Així doncs, necessitem definir una metodologia que avaluï el rendiment de les tècniques d'OF.

Aquesta tesi aporta la definició d'una metodologia que ens permet decidir quines parelles *"optical flow - mesura de confiança"* (OF-CM) estan millor preparades per a definir una cota de l'error de l'OF donat un nivell de confiança per a un DSS. Per tal de definir aquesta metodologia, la tesis engloba els següents punts:

- **Marcadors qualificatius.** Es presenten 3 gràfiques descriptives que avaluen de forma visual les capacitats de CM d'acotar l'error de l'OF. A més a més de les gràfiques descriptives, donada una gràfica representant la parella OF-CM, donem una qualificació automàtica que categoritza la gràfica donat el tipus de perfil.

- **Metodologia estadística.** Es proporciona una metodologia comparativa que permet determinar quina és la millor parella OF-CM per a acotar l'error de l'OF, aquesta metodologia consta de dues parts. Primer s'avalua el valor predictiu de la CM mitjançant la gràfica descriptiva. Després, per a una mostra de gràfiques descriptives calculades sobre unes seqüències de *training*, s'obté una corba genèrica que es podrà fer servir per a seqüències que no tenen *ground truth*. En el segon pas, s'avalua la corba genèrica obtinguda i les seves capacitats per a reflectir el valor predictiu de la mesura de confiança mitjançant ANOVA's.

La metodologia presentada mostra el potencial en aplicació clínica per a DSS. En concret, s'ha analitzat l'impacte de diferents artefactes en la imatge com ara soroll o deteriorament en el resultat final d'OF per a imatges del cor. També s'ha aplicat per a millorar la navegació dintre l'arbre bronquial en una broncoscòpia.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Image sequence analysis involves, among others, recognizing specific objects, computing their position, tracking them or determining motion for each point of the image. Sequence motion analysis of a sequence is a standout field used in a wide range of applications such as security aids (like detection of anomalous and unpredicted agents in urban scenes) [1], car driver assistance [2, 3], UAV guiding [4], 3D reconstruction [5, 6], occlusion detection [7, 8], background substraction [9, 10] or clinical support systems [11–15] among others. All these applications require the extraction of sequence motion as a first computational step to obtain the final output. In order that such motion can be effectively used in a decision support system, one needs that the system also decides whether the computed motion is reliable or not to avoid error propagation to the final outcome.

In the context of clinical decision support systems, reliable computation of motion and deformation is a key mandatory step. Two flagship examples of such clinical tasks are cardiovascular disease diagnose and navigation guidance during endoscopic exploration.

On one hand, dynamic functional disorders in the myocardium reflect most of cardiovascular diseases [16–18]. Thus, it is of prime importance an accurate visualization and computation of myocardium dynamics for its diagnose, treatment and follow up. A widely used technique to evaluate function damage is Tagged Magnetic Resonance (TMR). TMR sequences print a magnetic grid on cardiac tissue (see top left image in figure 1.1) which deforms along the cardiac cycle. In this manner TMR allows visualization of the intramural and wall motion of the myocardium (that is, internal tissue deformation). However, TMR images still lack from a sharp contrast between tagged pattern and tissue, in addition, such contrast decreases along sequence frames. In order to improve the reliability of the diagnose, physicians demand computer vision tools that help them to better analyze the image. Most applications analyze myocardium local deformation by computing the motion of each point of the sequence and local scores of diagnostic value, such as strain and torsion [19, 20].

On the other hand, intervention guiding has become a main issue in videobronchoscopy imaging. Videobronchoscopy is an endoscopic technique for the internal exploration of the respiratory pathway. This technique consists in visualizing the

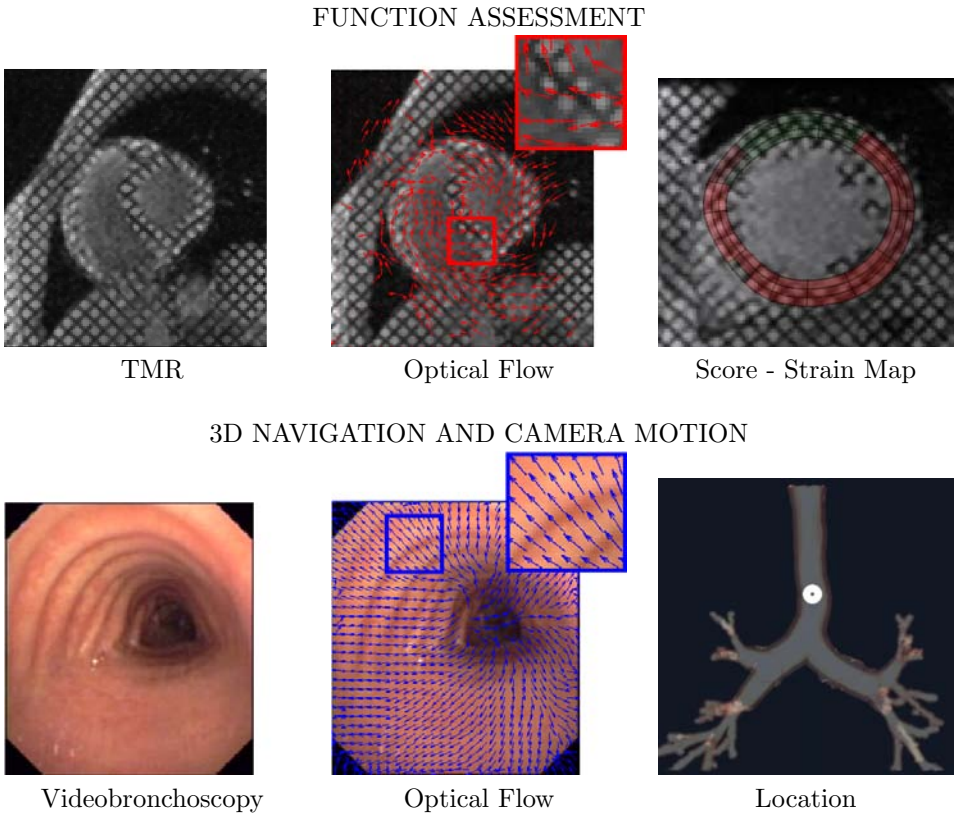


Figure 1.1: Medical applications requiring confident motion computation.

inside of pulmonary airways for diagnostic and therapeutic purposes [21]. The visualization is recorded using a bronchoscope that has a fiberoptic system that transmits an image from the tip of the instrument to an eyepiece or video camera at the opposite end. During diagnostic explorations the endoscopist must either reach the damaged tissue identified in a previous 3D scanner or perform on-line identification during the endoscopic exploration. In the first case, the bronchoscopist has to navigate through the bronchial tree until the cancerous nodule is reached. In the second case, it should visually identify and measure the scope and 3D side of the lesion. In both cases, camera lens distortion, projection artifacts and camera manual motion seriously hinder the observer performance that influences diagnoses yield regardless of their experience [22, 23]. Therefore, clinicians are in an urgent need of an online accurate computation of 3D measurements and location from analysis of videobronchoscopy explorations [24, 25].

Figure 1.1 illustrates the two clinical user cases requiring reliable computation of motion: cardiac function assessment on first row and bronchoscopy 3D guiding on second row. For cardiovascular disease diagnose we show, in the first column, a frame of Tagged Magnetic Resonance (TMR) showing the tagged pattern that allows track-

ing intra-wall tissue deformation along the cardiac cycle. The second column image shows the computed point-wise motion in red arrows and a close up of the computed motion. Third column shows an example of the optical flow application, in this case it is used to compute local scores of functionality disorders like the strain map. For 3D navigation through the trachea we show, in the first column, a videobronchoscopy frame. The middle image shows the optical flow computed over the sequence frame in blue arrows and a close up of the flow field. Third column shows an example of optical flow application which allows to know the location of the camera.

The two paradigms of clinical applications are illustrative of the two kinds of motion that can appear on sequence objects: rigid or elastic. Rigid motion is induced by the displacement of objects that have the same shape along a sequence. It is often present in natural scenes like car driving sequences and its principal application is object tracking or camera pose estimation. Rigid motion is commonly analyzed by applying techniques such as particle filters [26] or adaptive appearance models [27]. On the other hand, elastic motion is induced by the displacement of objects that may change its shape along a sequence. Elastic motion mainly arises in biomedical images (tissue deformation). A main feature of elastic deformation is that each pixel of an object has a different motion and thus, motion estimation has to be local. We note that for cardiac deformation motion vectors, although continuous, do not have the common trend shown in the bronchoscopy navigation, whose motion vectors have the typical radial disposition arising from camera central motion. Figure 1.1 shows an example of rigid and elastic motion for the two clinical cases of use. Second row shows an example of global rigid motion due to camera navigation in videobronchoscopy, whereas the first row illustrates the local elastic deformation of myocardial tissue observed in TMR sequences. The milestone for computing point wise local estimation of motion is Optical Flow (OF), introduced by Gibson et al. [28, 29].

Optical Flow is the pattern of apparent motion of objects, surfaces, and edges in a visual scene caused by the relative motion between an observer (eye or camera) and the scene [30]. That is, given a sequence, OF is the projection of the 3D scene motion into the image and is given by a dense vector field that indicates for each pixel of a frame, its displacement to the next frame.

The core of OF computation is the assumption that an image local feature keeps constant along the sequence. In most cases such feature is the intensity and it is referred as Brightness Constancy Constraint, BCC. The BCC assumption provides one equation and two unknowns, thus, the original OF is an ill-posed problem (known as aperture problem) that requires further information on the motion vector in order to be solved [31,32]. The first attempt to solve the OF equation are local techniques [33, 34], which assume spatial coherence (i.e. points move like their neighbors) and solve the aperture problem by deriving a local system of equations from the OF equations in a neighborhood of each pixel. The main concern about local techniques is that the system of equations is not solvable for all pixels of the image and thus they do not provide dense motion fields. In addition, due to the spatial coherence assumption, the computation of optical flow is not reliable in motion boundaries. The second approach to solve the OF equation are variational schemes introduced by Horn and Schunck [35]. In contrast to local techniques, variational approaches are able to compute dense motion fields [35–40]. They compute motion by finding the minimum of an energy

functional [35] which combines two terms: the data-term and the smoothness-term. The data-term puts into correspondence one frame with the following one, whereas the smoothness-term determines the global regularity properties of the vector field across the image [41]. Even though the smoothness term is the key to obtain a dense flow field, in some areas the vector field might be over-regularized (especially at motion boundaries). Figure 1.2 illustrates the impact of the regularity term on the OF sharpness and quality. We show the OF computed using two variational techniques with different OF regularization terms: an L^2 norm method, in this case the classical Horn and Schunck [35] and a L^1 total variation method labelled classic-NL [40]. Results are obtained for two representative sequences of two benchmark synthetic databases with ground truth: Middlebury[1] and Sintel[2]. Both sequences have independent moving objects, which represent a main modelling challenge for the data terms. Motion vectors are shown in color code images, where for each pixel, color indicates the direction and hue the magnitude of the optical flow vector. We have selected two highlighted in the upper close-ups. Such areas show the impact of the regularity term, even though in the case of L^1 norm the motion boundaries are sharpener, both optical flow techniques fail in motion boundaries and occlusions.

Although there has been an increasing interest in developing new techniques to obtain dense flow fields, minimize over-regularization and keep motion discontinuities [36, 40, 42–45], the OF formulation still can not provide reliable computations in the sequence domain due to input data errors, propagation errors, illumination changes, etc. Thus, in the context of decision support systems that use optical flow, it is of prime importance the definition of a framework and a general road map for its performance evaluation [46]. Such framework requires discarding those regions where OF is neither reliable nor accurate. The most common way of measuring OF accuracy is by computing its deviation from the true motion vector. This suffices to quantify the overall performance, but it is useless at locating areas of poor performance in real-time applications where no ground-truth is given. In real situations, the absence of ground truth forces to assess OF quality using quantities computed from either sequences or the computed optical flow itself. These quantities are generally known as Confidence Measures, CM.

Confidence measures should be an indicator of the accuracy of the output of an optical flow algorithm obtained from the analysis of input or output data. 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 [47].

In order that a measure is useful for bounding errors, the plot between the measure and endpoint errors should show a monotonic tendency. That is, high errors always

[1] http://vision.middlebury.edu/flow/eval/
[2] http://sintel.is.tue.mpg.de/results

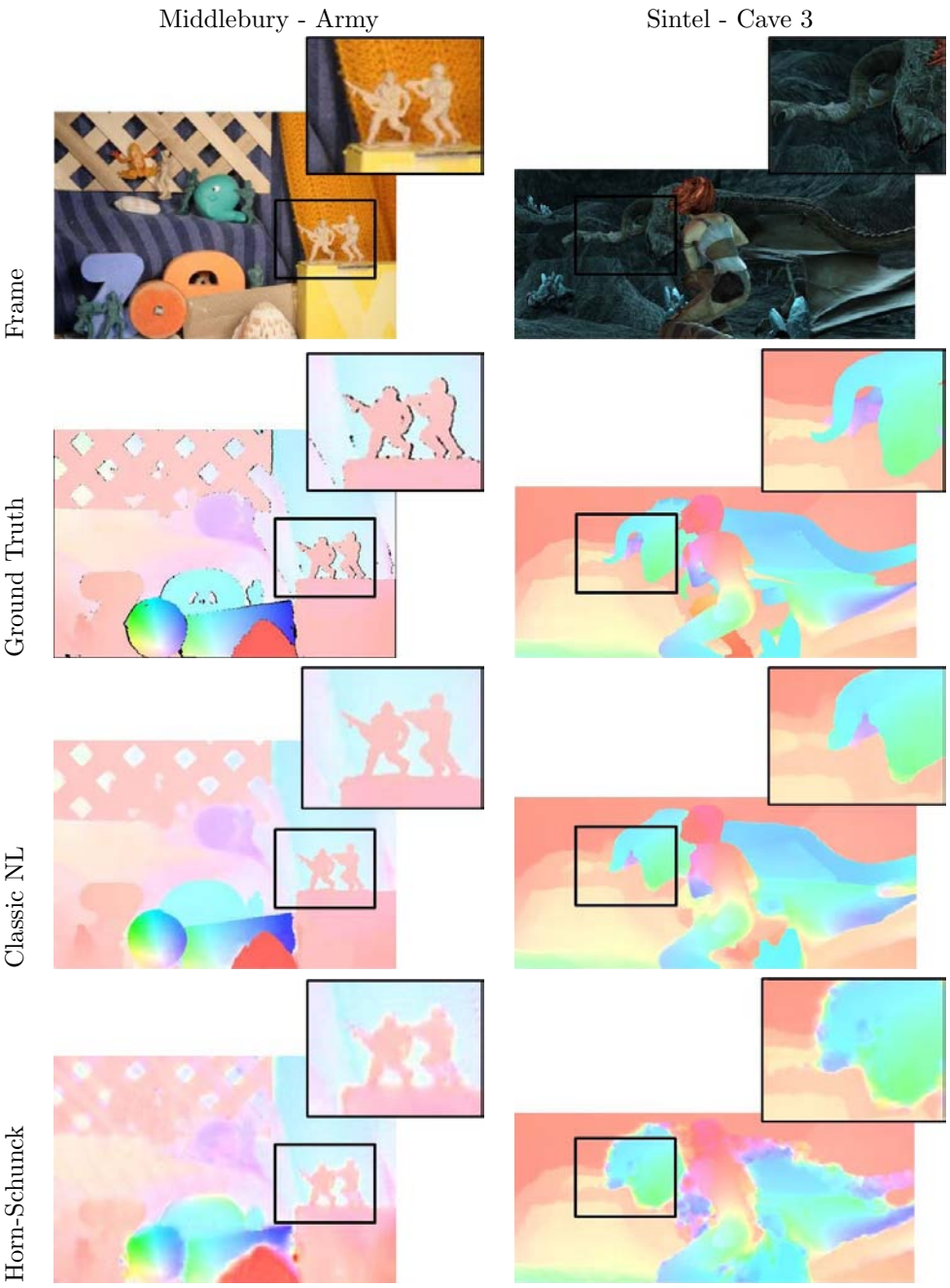


Figure 1.2: Comparison between ground truth and computed OF.

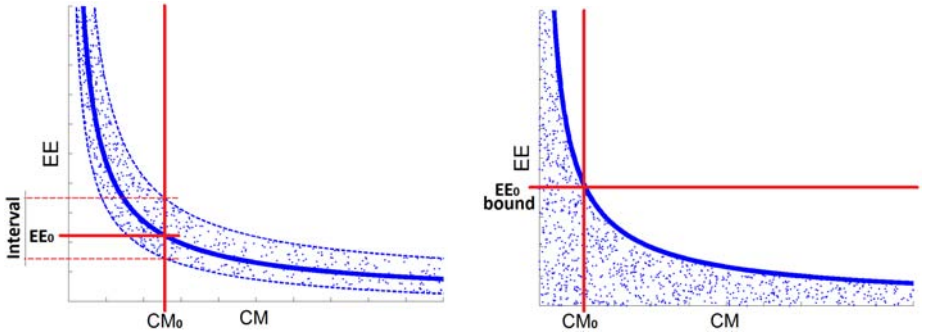


Figure 1.3: Difference between error prediction (a) and error bounding (b).

correspond to either high or low CM values. In this work, we will assume that high OF errors correspond to low CM values, so that scatter plots of OF error and CM values have a decreasing profile.

Plots in figure 1.3 illustrate the difference between the concepts of error prediction and error bounding. We show the scatter plot of the confidence measure values (CM) versus the OF End-Point Error (EE) [48]. The x-axis corresponds to CM values (normalized in range $[0, 1]$) and y-axis corresponds to OF End-Point Error (EE). In the first case, shown in fig. 1.3 (a), there is a clear functional correlation between CM and EE and, thus, it could make sense fitting a regression curve by least squares (solid blue line). In this context the fitting error which corresponds to the scatter along the curve determines the quality of the error prediction given by the measure values. Such scatter is visualized in fig. 1.3 (a) by the vertical interval around each EE value, $EE = EE_0$ indicated with a red horizontal line. A functional link between CM and EE is never met in practice and the most usual case is that CM-EE plots are pure scatter plots like the one shown in fig. 1.3 (b). In this second case, we do not have any functional dependency but a scatter plot following a decreasing pattern, which upper bound for each CM values (solid blue line) can also be used for error bound. In this case, a given value of CM, CM_0, 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.

Even if we have a proper confidence measure we still need a way to evaluate its capabilities to discard pixels with an OF prone to have large errors. In addition, each confidence measure is more capable to bound an specific type of OF error, so we also need to determine which confidence measure performs better with an specific optical flow method. Therefore, it is of prime importance the definition of a framework and a general road map for the evaluation of optical flow and confidence measures performance, as discussed in [46].

1.1 Thesis goal

The main goal of this thesis is to design a framework able to decide which pairs *"optical flow - confidence measure"* (OF-CM) are best suited for optical flow error bounding given a confidence level determined by a decision support system. In order to achieve the goal, this thesis covers the following points:

- **Descriptive Scores for CM Performance Evaluation.** First we design the descriptive plots and scores that describe the profile of CM-EE scatter plots and quantify its decreasing tendency.
- **Statistical Framework for Comparison of CM Predictive Value.** Second, we design a statistically based analysis to compare the different OF-CM pairs in order to choose the best suited to bound the error for a given application.

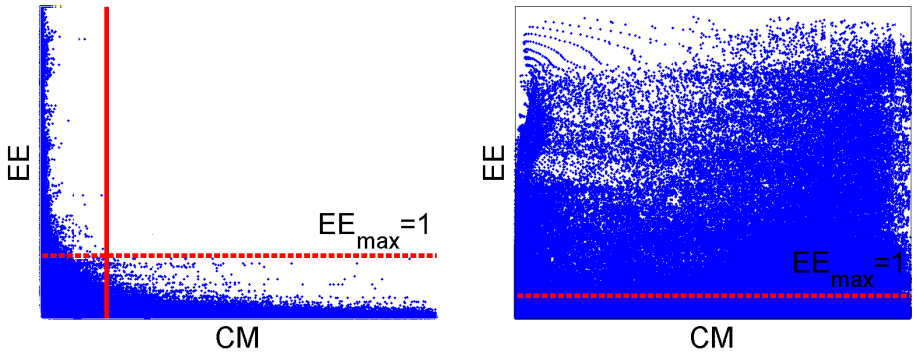
1.2 State of the Art and Contributions

This section summarizes current approaches on confidence measures and its performance evaluation in order to state our contributions to each of the above points.

In order to use optical flow in a decision support system, a mechanism to detect sequence pixels that have high error in their computations is of prime importance. In this context, Confidence Measures (CM) should be an indicator of the accuracy of the output of an optical flow algorithm. It should be noted that a confidence measure can provide at most an upper bound of OF error at each pixel, not its real value (according to numerical error analysis [47]). This implies that high values of the confidence measure should ensure a low OF error, while for low CM values errors could take any value. Points that have high error and high value of the confidence measure are unpredictable points which CM can not discard and, thus, should be the least possible.

Figure 1.4 helps to understand this idea. It shows the scatter plot of the confidence measure values (CM) against the OF End-Point Error (EE) [48]. The x-axis corresponds to CM values (normalized in range $[0, 1]$) and y-axis corresponds to OF End-Point Error (EE). We plot an horizontal line at $EE = EE_{max} = 1$ representing the maximum error allowed by the application for a better comparison. Figure 1.4(a) presents a clear decreasing profile which allows determining the minimum CM value that guarantees that points with CM above such threshold (vertical line) have an EE below 1. Meanwhile the uniform distribution of CM-EE scatter in figure 1.4(b) makes impossible the definition of such threshold on CM.

Confidence measures can be formulated from either an analytic or a probabilistic point of view. Analytic approaches either use the energy [49,50] or the image structure (gradient magnitude [31], structure tensor [51]) 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 [52] or its variability with respect perturbations in the model [53]. Probabilistic approaches are more flexible and not



(a) CM able to bound OF error (b) CM can not bound OF error

Figure 1.4: Scatter plot confidence measure (CM) against optical flow error (EE).

necessarily linked to any source of error. Furthermore, they can even be used to get a confidence fusing all previous measures [54] and, thus, relate to other OF error sources (see chapter 3).

In their seminal work on optical flow evaluation, Barron et al. [31] emphasize the importance of confidence measures to examine optical flow methods. They also carry out a first comparison by applying the confidence measures and, after thresholding by the CM on the sequences, visually compare the performance of the OF methods. Following the same method, Liu et al. [55] visually compare two different CM and Fazekas et al. [56] study the efficiencies of other two CM in four OF methods. Some years later, Bruhn et al. [38] visually check that the pixels suppressed by the confidence measure are indeed those pixels having the lowest performance of the OF. That is, they visually correlate the results of the CM selection with the accuracy of OF. Although these results have been the first steps towards a comparison of confidence measures within a single framework, they only provide visual comparisons.

An early general attempt to define a type of CM quantitative evaluation for simultaneous comparison has been made by Bruhn et al. [49]. They validate the quality of CM by means of Sparsification Plots (SP). 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, without taking into account the variability. Under the assumption that higher values of CM are associated to lower flow errors, SP should have decreasing profiles. An increase in their values for the higher removed fractions indicates artifacts in the decreasing dependency possibly due to a high error despite a high CM. However, the inverse does not always hold and random uniform dependencies could produce sensible plots.

Existing descriptive plots can only provide a visual assessment of the performance of CM that is not enough for standardized quantification and comparison. An alternative is to compute a single overall score for each CM, like the average EE across test sequences proposed in [54]. Although this is a compact way of comparison, a global score might not suffice to detect significant differences across methodologies [57]. An-

other common concern is the impossibility to explore sources of variability in OF performance and select which CM is best suited for a given OF. Finally, comparison results should generalize to sequences with similar experimental settings than the ones used at the evaluation stage with a given confidence.

1.2.1 Contributions

This PhD thesis contributes to reliable computation of optical flow for improved decision support systems in the following aspects:

- **Descriptive scores.** As a first step, we summarize and analyze the sources of inaccuracies in the output of optical flow algorithms. Second, as we have seen in the state of the art section, in the literature there are no robust and descriptive tools to represent the performance of confidence measures according to the optical flow error. We present several descriptive plots that visually assess CM capabilities for OF error bounding in terms of the conditional distribution function measuring the probability that CM can not bound the error. In addition to descriptive plots, given a plot representing OF-CM capabilities to bound the error, we provide a numeric score that categorizes the plot according to its decreasing profile, that is, a score assessing CM performance.

- **Statistical framework.** We provide a comparison framework that assesses the best suited OF-CM pair for error bounding that uses a two stage cascade process. First of all we assess the predictive value of the confidence measures by means of the Sparse-Density Plots. Then, for a sample of SDP plots computed over training frames, we obtain a generic curve that will be used for sequences with no ground truth. As a second step, we evaluate the obtained general curve and its capabilities to really reflect the predictive value of a confidence measure using the variability across train frames. Furthermore we apply ANOVA to decide which pairs OF-CM perform better and better assess the optical flow error.

Figure 1.5 illustrates the thesis contributions. The top image shows the current optical flow situation, and the bottom one summarizes the thesis contributions.

The presented framework has been applied to improve clinical decision support systems. In particular, it has been applied to improve the following two systems:

- **Cardiac Diagnose Systems.** The presented framework has been applied to a synthetic cardiac database to analyze the impact of the different image artifacts such as noise and decay to the output of optical flow.

- **Bronchoscopy Guidance.** We have used the presented framework in manually annotated dataset of bronchoscopy guidance to improve the navigation inside the bronchial tree.

Figure 1.6 shows an scheme of the application of the thesis contributions. Applying the presented statistical framework to a selected application of optical flow, it provides a more reliable input for the application, and an improvement and reliability of the final results.

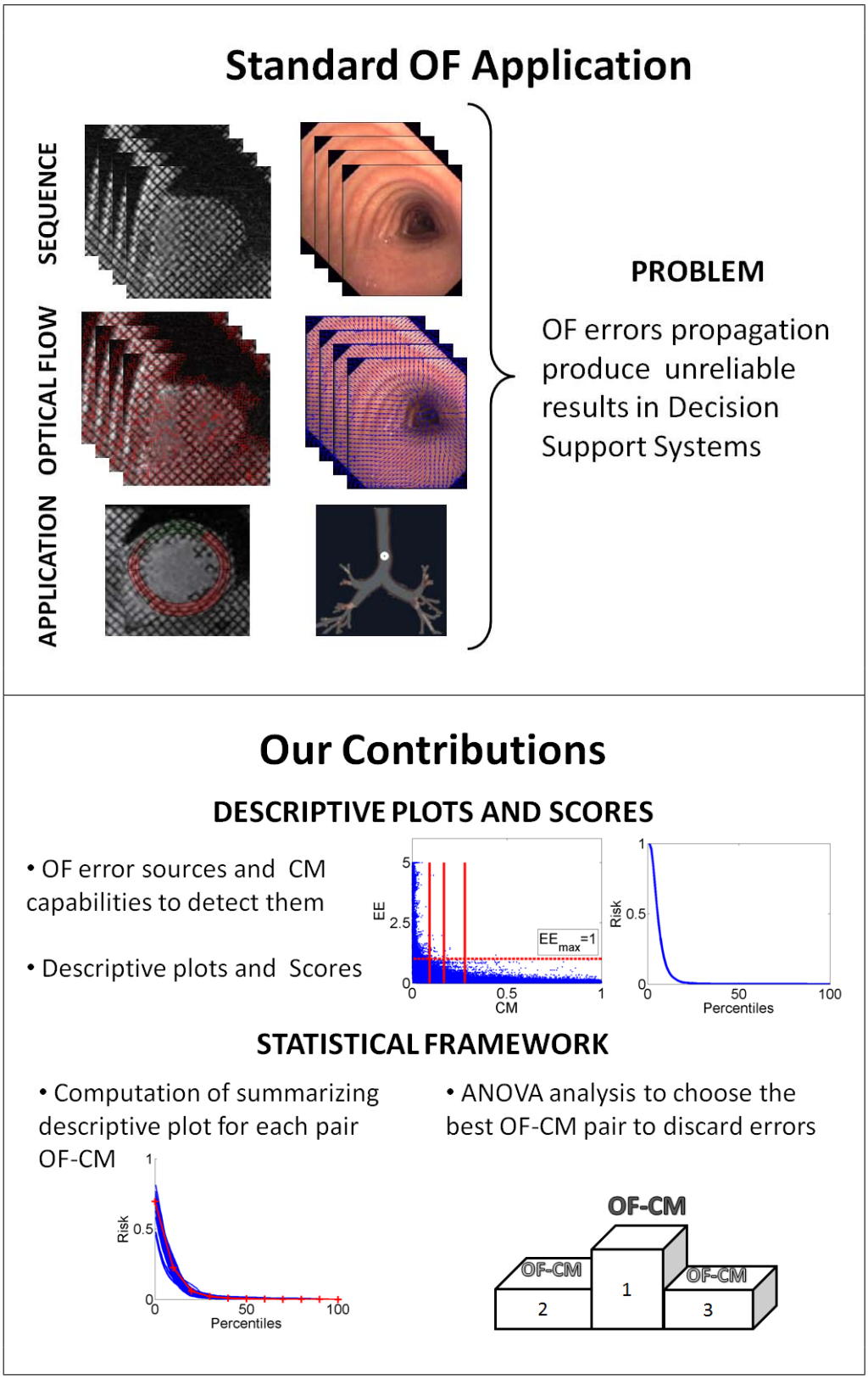


Figure 1.5: Thesis contributions.

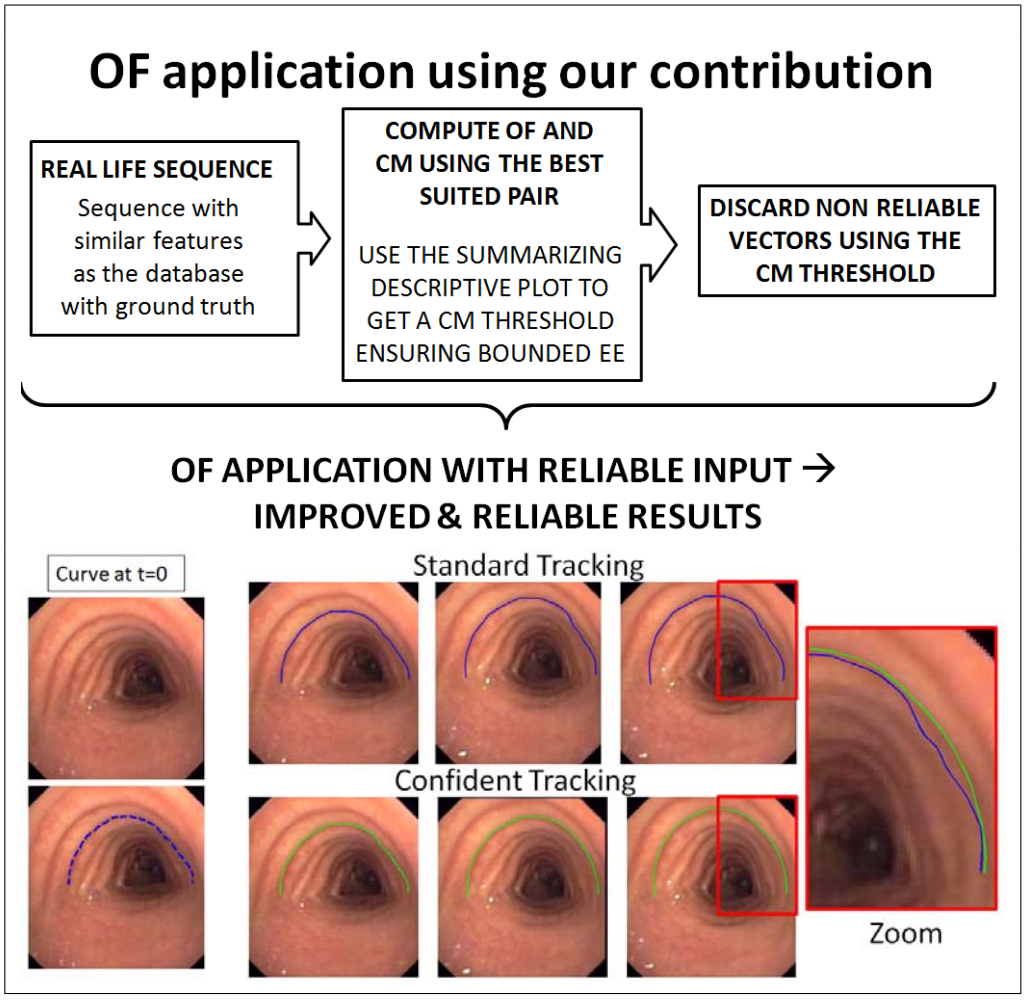


Figure 1.6: Application of the thesis contributions.

The remains of the thesis are organized as follows: chapter 2 compiles a review of the different optical flow techniques, confidence measures, and current approaches for performance evaluation of confidence measures. It also summarizes the different statistical approaches required to define the statistical framework. Chapter 3 analyzes the optical flow error sources and the capabilities of the state of the art confidence measures to bound the different optical flow error sources. In addition, in this chapter we define the descriptive plots and scores to evaluate the performance of the confidence measures. Chapter 4 defines a statistical framework able to choose the best pair OF-CM able to properly bound the optical flow error for a given decision support system. Chapter 5 applies the defined statistical framework to improve clinical decision support systems, in particular, cardiac diagnose systems and bronchoscopy guidance. Finally, chapter 6 outlines the main conclusions of the thesis and future lines of research.

Chapter 2

Theoretical tools and Databases

Optical Flow (OF) techniques are a widespread tool for computing pixel-wise motion between consecutive frames. In this chapter, we summarize how optical flow is computed and also the state of the art techniques. In addition we summarize the different current tools that are used to determine the accuracy of optical flow methods.

Besides, the statistical framework presented in this thesis analyzing the performance of different pairs (optical flow technique, confidence measure) requires several statistical tools, which are also presented in this chapter.

2.1 Basics of Optical Flow

OF is defined as the velocity vector field that transforms one frame into the following one. It assumes that object appearance (given by image intensity) keeps constant along sequence frames (Brightness Constancy Constraint, BCC). Under such assumption, the vector field given by OF puts into correspondence pixels in consecutive frames that have the same appearance (intensity). In mathematical terms, these requirements are formulated as follows: let $I(x(t), y(t), t)$ (I for short) be a sequence and $(x(t_0), y(t_0), t_0)$ an image pixel at time t_0. If we assume the BCC, the equation to solve is the following:

$$I(x(t+t_0), y(t+t_0), t+t_0) = I(x(t_0), y(t_0), t_0) \tag{2.1}$$

Considering that displacements are small and the image sequence varies smoothly along the spatial and temporal coordinates, we can use first-order Taylor expansion in time t_0, obtaining:

$$I_x x_t + I_y y_t + I_t = 0 \tag{2.2}$$

Since, OF is a vector field, we can define $W = (x_t, y_t) = (u, v)$ and re-write the equation (2.2) into the following compact form:

$$< \nabla I, W > + I_t = 0 \tag{2.3}$$

The above equation will be called OF equation.

The solution of (2.3) gives, for each pixel of the image, the estimation of motion given by OF. The application of OF formulation to real life sequences presents three weighty limitations: BCC is not always fulfilled, sequences should have small temporal deformations and the equation to solve the optical flow is an ill-posed equation.

In real life sequences, BCC is not always fulfilled due to illumination changes, physical properties or image acquisition devices. Thus, the computed OF may not correspond to the real motion. One way of ensuring the BCC is to either change the feature to keep constant along sequence frames or the representation space of the sequence [31, 49]. There are two kinds of alternative features, region descriptors and edges and corners descriptors. Region-based techniques [50, 58] match shift region descriptors, while feature-based approaches [59–61] seek correspondences of characteristic image features such as edges or corners. Besides, methods that change the image representation space replace brightness by a filter response. In this case, the velocity vector field is defined from the phase behavior of band-pass filter outputs in the Fourier domain and so they are called phase-based approaches [34, 62, 63].

The second limitation of the OF formulation is that, independently of the feature to keep constant, OF equation is based on derivatives. Therefore, motion is not properly recovered for large deformations and a high temporal resolution sequence is needed. To avoid such problems there are wrapping techniques.

Finally, OF computation is an ill-posed problem since equation (2.3) introduces one constraint with two unknowns so it can not be uniquely solved. Indeed equation (2.3) can only recover motion along the image gradient (normal to the image level sets). That is, if we express $W = \omega_1 \nabla I + \omega_2 \nabla I^{\top}$ and develop the scalar product, we have:

$$< \nabla I, W >= \cos\theta \cdot \|W\| \cdot \|\nabla I\| = \omega_1 \cdot \|\nabla I\| \tag{2.4}$$

for θ the angle between the motion vector and ∇I. By replacing (2.4) in (2.3) we obtain:

$$\omega_1 \cdot \|\nabla I\| + I_t = 0 \Leftrightarrow \omega_1 = \frac{-I_t}{\|\nabla I\|} \tag{2.5}$$

Figure 2.1 graphically shows the projection of W over ∇I.

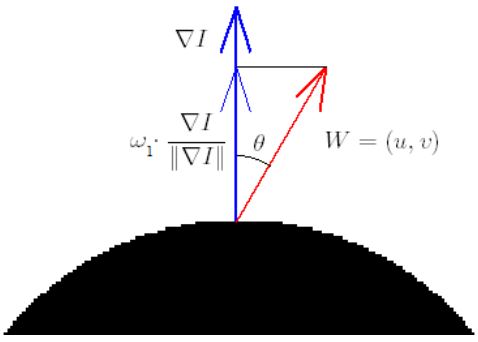


Figure 2.1: Geometric interpretation of the OF.

This is a phenomenon called the aperture problem and arises as a consequence of converting a two dimensional problem into a one dimensional one. Depending on the geometry of the object and the kind of motion, OF equation can properly recover the motion or not. That is, in points where motion is perpendicular to the image level set, OF equation recovers the whole motion, while in points where motion is tangent to the image contours, the OF equation does not recover motion at all. The remaining possible motions in a point will be partially recovered. Figure 2.2 shows three different scenarios. Blue arrows correspond to the motion that OF equation recovers while red arrows correspond to the real motion of the object. In figure 2.2(a) motion is perpendicular to the edge; therefore, blue arrows coincide with the red ones, that is, we can recover the motion. In figure 2.2(b) motion is tangent to the edge, consequently no motion is recovered. Finally, in figure 2.2(c) motion is oblique to the edges, thus we will recover the normal component of the motion. Notice that, since in the corner there are two different equations of the OF, motion can be properly recovered.

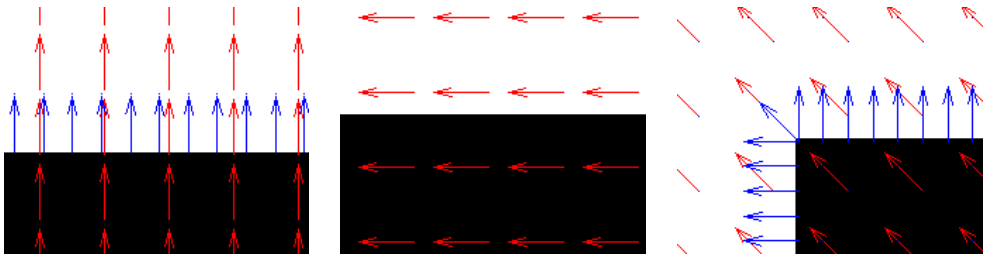


Figure 2.2: Consequences of the aperture problem. In red the motion of the object, and in blue the solution given by the OF.

In order to minimize the aperture problem some outstanding local techniques propose an equation system by assuming some properties of the vector field [33,34,64]. Lucas and Kanade [33] is a differential approach that assumes that motion is constant in a local neighborhood around each pixel and then, applying least squares, computes the solution. Fleet and Jepson [34] is a phase-based technique that defines the velocity in terms of the gradient of the phase output of a Gabor filter, and thus the vector is the least squares solution of an equation system. Another representative method to estimate motion locally, specifically defined for tMRI images, is the HARmonic Phase (HARP) [64,65]. HARP method tracks the phase of the Fourier coefficients associated to the tagged pattern, and then uses a Newton-Raphson approach for seeking the points that keep constant phase values between consecutive HARP images. Since the correspondence is computed separately for each pixel, the resulting vector field is often irregular.

Local techniques solve the aperture problem in some pixels but do not produce dense flow fields. In contrast, variational techniques, developed for the first time by Horn and Schunck [35], produce dense flow fields by combining into a variational framework a data-term (which assumes constancy in the object appearance) and a smoothness-term (which models the behavior of the flow across the image). These

approaches compute the OF (W) by finding the minimum of the following energy functional:

$$\varepsilon(W) = \iint \underbrace{\|f(\nabla I, W)\|^2_{L^2}}_{Data-Term} \, \mathrm{d}x \, \mathrm{d}y + \alpha \iint \underbrace{g(W, \nabla W, ...)}_{Smoothness-Term} \, \mathrm{d}x \, \mathrm{d}y \tag{2.6}$$

where f is the data-term, g the smoothness-term, α is a constant regularization parameter, $\|\cdot\|_{L^2}$ is the L^2 norm and ... denotes higher order terms.

The data-term is a function that puts into correspondence one frame with the following one. Existing approaches use either the OF equation [35, 36] or the system equation provided by local techniques such as Lucas and Kanade [38] or Fleet and Jepson [63]. OF equation as data-term still presents the aperture problem, so it needs to be solved in the variational framework together with the smoothness-term. In the case of local techniques, since data-terms are own-solvable, the smoothness-term is introduced to regularize the velocity vector field. Thus, they are combining the robustness of local methods and the density of the variational ones.

The smoothness-term determines the global properties of the vector field [41]. The first approaches defined the smoothness-term in the L^2 space to ensure differentiability [35, 36, 38]. The main limitation is that the solution can be over-regularized in cases that there are occlusions or discontinuities in the velocity vector field. In order to overcome that, total variational methods define g in the L^1 space [38, 40, 66, 67]. The problem of those techniques is that there are functions from different spaces in the same variational framework, which is mathematically inconsistent and also there is no robust theory which assures the reliability of the solution.

In addition to the data and smoothness terms, the parameter α provides a trade-off between both terms. The bigger the α is, the smoother the flow field becomes. In most cases, this parameter is constant and can be learned from a training set [40]. Other works [36, 63] consider that the smoothness-term should play an important role in those points where the data-term does not provide motion information. For instance, Nagel and Enkelmann [36] weight the variational by means of $\|\nabla I\|$. The main idea is that $\nabla I = 0$ indicates a flat region, since in these points we can not recover any motion, the scheme gives more weight to the smoothness-term. However ∇I may not reflect all regions where the data-term can not properly recover motion. In this context we claim that α has to reflect the theoretical conditions that assure a good solution of the data-term. In this fashion, [39] introduces the amplitude of the response of Gabor Filters in the weights with an evident improvement on the computation of the OF. This approach has been applied to medical images, in particular, to the assessment of the left ventricle motion.

2.1.1 Accuracy score of optical flow and ground truth databases

In order to quantify the accuracy of the optical flow, we need to evaluate the difference between the computed optical flow and the ground truth. In the literature we can find several ways to compute the accuracy [48], but the most extended way to measure it is the End-Point Error (EE) [48]:

$$EE = \|W_C - W_{GT}\|_2 \tag{2.7}$$

where W_C is the computed OF and w_{GT} is the ground-truth of the OF.

The definition of ground truth for optical flow in real life scenes is a challenging task. There may be inherent errors on the ground truth due to sequence properties such as occlusions or noise, or discretization errors induced by the generation of the ground truth itself. This problems for the ground truth also can be found in synthetic databases in a lower level. In the literature we can find several optical flow databases [48,68–70], but, since all optical flow methods are somehow trained to outperform for such databases those ones have a limited time. In this thesis we use the following databases:

- **Middlebury [48].** This database contains real-life and synthetic sequences with ground-truth. The sequences contain several independently moving objects, thin structures, shadows and foreground-background transitions. It contains displacements up to 20 pixels per frame, and the sequences have up to eight frames.
- **Sintel [68].** This database is derived from the open source 3D animated short film Sintel, thus the authors refer to it as a semi-realistic database. It contains sequences with large motion displacements, specular reflection, motion blur, defocus blur and atmospheric effects and, thus, it covers a complete bunch of sequence features. In this database we can find displacements up to 40 pixels per frame, and the sequences contain up to 40 frames.
- **Cardiac database [71].** This is a synthetic database simulating MR images, and uses the cardiac motion simulator defined by Arts and Waks [72, 73]. The dataset consists of five short axis slices sampled across the prolate sphere. Every slice has 50×50 pixels. The spatial period of the tagging patterns is set to 6.6 pixels. Rician noise is added with a constant SNR of 25 over time, defined as $SNR = \frac{\mu}{\sigma}$ with μ the mean signal and σ the standard deviation of the noise [74]. Two data variants have been generated: images with noise and decay and clean data without noise and decay.
- **Bronchoscopy database.** This database consists of videobronchoscopy explorations of the tracheal structures, all the images are courtesy of Hospital de Bellvitge in Barcelona, Spain. We have selected a sample of representative sequences and for each sequence we have selected a ring to track. Such rings have been manually labelled as ground truth.

2.2 Confidence Measures

In order to use optical flow in a confident decision support system, a mechanism to detect sequence pixels that have high error in their computations is of prime importance. In this context, Confidence Measures (CM) should be an indicator of the accuracy of the output of an optical flow algorithm. It should be noted that a confidence measure can provide at most an upper bound of OF error at each pixel, not its real value (according to numerical error analysis [47]). This implies that high values of the confidence measure should ensure a low OF error, while for low CM values errors

could take any value. Points that have high error and high value of the confidence measure are unpredictable points which CM can not discard and, thus, should be the least possible. Therefore, the final purpose of a confidence measure should be the detection of numerical errors and also, provide an upper bound for the flow error.

Confidence measures can be formulated from either an analytic or a probabilistic point of view. Analytic approaches either use the image structure (gradient magnitude [31] or structure tensor [51]) or the energy [49, 50] as indicators of confidence. Probabilistic approaches define confidence in terms of probabilistic distributions from either the flow field itself [52] or its variability with respect perturbations in the model [53]. Ahoda et al. [54] use supervised learning to estimate a confidence value for each pixel of the image.

Analytic

Since the data-term is formulated using the image partial derivatives, the first measures were defined in terms of the image local structure [31, 51, 75]. One of them is the determinant of the matrix of the structure tensor of the image, we will refer to it as C_d. Since the data-term is formulated using the image partial derivatives, several measures are defined in terms of the image local structure, either gradient or structure tensor. Gradient-based measures [31] is defined as the magnitude of the gradient. Note that for large values of the magnitude of the gradient we expect reliable motion vectors. However, large gradients usually denote occlusions or noise [49], and in those pixels the optical flow computation is not reliable. In order to minimize the impact of noise, structure tensor based measures use information about the local structure of the image. These measures are especially well suited for LK-based schemes. There are several measures derived from the structure tensor: determinant [31], trace [75], lowest eigenvalue [51] among others. We will consider the determinant-based measure and we will refer to it as C_d.

Some of them are based on the principle that the OF equation and LK system require some image properties in order to make sense. In this context, gradient-based measures are defined for OF equation and structure tensor ones for LK.

The energy-based measure takes into account that variational techniques compute optical flow by minimizing an energy functional (2.6), and thus, the confidence measure is computed evaluating the flow field over the functional. Under the (sensible) assumption that all constrains have been taken into account in the definition of the functional, the computed flow field will be accurate in the measure that its local energy is low. Meanwhile for pixels where such energy is high, the flow field does not fulfil the model and, thus, might have a higher error. We will refer to this confidence measure as C_e and it is defined in [49]. Under these considerations, the following measure is proposed:

$$C_e = \frac{1}{D(u, v, \nabla I) + \alpha S(\nabla u, \nabla v, \nabla I) + \epsilon^2} \tag{2.8}$$

where ϵ prevents dividing by zero.

Probabilistic

Measures based on pattern analysis of computed flows are an alternative for defining confidence measures regardless of the model assumptions [52]. In many applications, flow fields follow similar local motion patterns. If such motion patterns are learned a priori, then a classifier can be used to define a confidence measure. The measure introduced in [52], which we note as C_s, derives natural motion statistics from sample data and carries out a hypothesis test to obtain confidence values for the computed flow. The method depends only on the resulting flow field and on the prior knowledge learned from a database. The confidence measure 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. Since the formulation is not straight forward, we refer the reader to the paper [52] for more details.

The measure defined in [53] quantifies the uncertainty of the flow method, that is, in those points where the flow field varies, the computation is not reliable. They compute such measure using bootstrap resampling. We refer to this measure as c_b, and we will consider the inverse of ψ_{bootg} defined in [53] eq.(15).In order to have a decreasing dependency with the accuracy we consider the inverse of the value:

$$c_b = \frac{1}{\psi_{bootg} + \epsilon^2} \tag{2.9}$$

for ψ_{bootg} defined in [53] eq.(15).

A different attempt that uses confidence measures to improve OF computation is presented in [54]. They use supervised learning to estimate for each pixel a confidence value for the computed flow vectors. They estimate if a flow algorithm is likely to fail in a specific region. Then, they can combine the output of several flow fields and combine them selecting for each pixel the one that performs best.

2.3 Confidence measure performance evaluation

The most extended way to represent the performance of confidence measures and its link to flow error is by means of the sparsification plots [49]. Such plots are given by the remaining mean error for fractions of removed flow vector having increasing confidence measure values (CM). That is, CM is sorted in increasing order, the $n\%$ of the flow vectors having low value of the measure are removed, and finally the average error of the remaining vectors is computed. The scatter plots in fig. 2.3 illustrate the computation of sparsification plots for two representative cases selected from the Middlebury database. For a given removed percentage (vertical line in scatter plots and x-axis in sparsification plot below), arrows indicate the points that are considered for the computation of average errors (y-axis in sparsification plot).

Under the assumption that higher values of CM are associated to lower optical flow errors, sparsification plots should have decreasing profiles. An increase in their values for the higher removed fractions indicates artifacts in the decreasing dependency possibly due to a high error despite a high CM. However, the inverse does not always hold and random uniform dependencies could produce sensible plots. This is the

case of the second representative sequence shown in fig. 2.3. Even if the dependency shown in the scatter plot is worse in the first sequence, its sparsification plot (blue line) indicates a better performance for high fractions.

Besides a poor power for assessing decreasing dependencies between confidence measures and flow error, sparsification plots are unable to properly detect if a measure is appropriate for giving a bound on optical flow accuracy. This is mainly due to the fact that its computation only considers confidence measure values for removing pixels regardless of optical flow error. Therefore, if the distribution of errors for high CM values concentrates around zero, the sparsification plot will be low even if we have some outliers with high errors. It follows that the distribution of, both, flow error and CM should be taken into account in order to properly measure the capability of confidence measures for bounding the error.

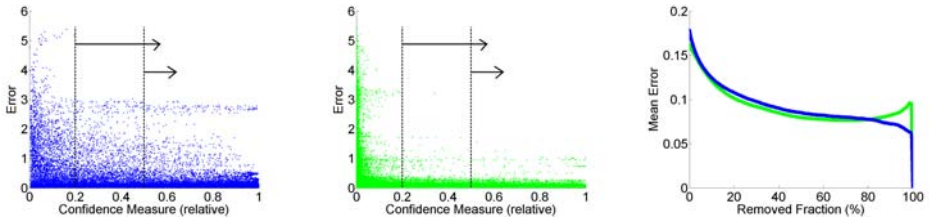


Figure 2.3: Sparsification plots. On the left and on the middle the scatter plot of a CM and the error. On the left a poor CM, on the middle a good one. The sparsification plot of both measures on the right.

2.4 Statistical Tools

In order to analyze the performance of different pairs OF-CM over a sequence, for every two consecutive frames, and decide which pair OF-CM performs better, we need information about:

- Performance variability: we want to know how much do the results vary depending on the sequence, and also how significant is the difference between the performances of the different pairs OF-CM.
- Effect of factors of interest: we wonder which is the impact of a certain factor on the performance.
- Generalization of the results: we need to know if a pair OF-CM can perform similarly in different conditions, in this case with sequences with different features.

For that, we present the following statistical tools:

- Confidence intervals
- Hypothesis tests
- T-test and Analysis of Variance (ANOVA)

2.4.1 Confidence Intervals

Given a sample population, a confidence interval provides the estimated range of values that includes an unknown population parameter, which is noted as ξ.

In order to compute a confidence interval we need a confidence level, an statistic and a margin of error. The confidence level, α, is the probability that the interval produced by the method employed includes the true value of the parameter ξ. Since the statistics associated to the sample are unknown, the most usual statistic is the mean of the population, μ. Finally, the range of confidence interval is defined by the *sample statistic* $\pm$ *margin of error*.

In our framework the confidence interval is a useful tool to define a general curve computed over a sample population of curves. The purpose of that generalized curve is to apply it for sequences without ground truth where is not possible to compute the SDP curve. Thus, due to the properties of the confidence interval we can assure for a given probability that such curve is properly defined.

2.4.2 Hypothesis Test

In case of hypothesis tests, we do not want to estimate a parameter of a given population but to decide between two alternatives of its value. For instance, decide between the alternatives of the population mean $\mu = \mu_0$ and $\mu \neq \mu_0$.

The hypothesis tests follow the same concept as *someone is innocent until proven guilty*. To perform the tests, two hypothesis are defined, the null hypothesis, H_0, and the alternative one, H_1. In our example innocent would be the null hypothesis, and guilty the alternative one. On one hand, the null hypothesis (H_0) is the assumption to be contrasted. Thus, it is considered that H_0 is true, and this hypothesis can or can not be rejected. On the other hand, the alternative hypothesis is the opposite to the null one, and it is usually the hypothesis that we want to proof. This hypothesis can or can not be accepted. If we can reject H_0, then the alternative hypothesis (H_1) becomes accepted, that is, we accept that the sample gives reasonable evidence to support the alternative hypothesis for the given confidence level. It is important to note that not rejecting H_0 does not mean that H_1 is false, but we have not enough evidence to accept one or the other. In that case, the confidence interval and the p-value indicates which hypothesis is more probable. The p-value is the estimated probability of rejecting the null hypothesis. So, what is the difference between the significance level and the p-value? Whereas the significance level α is a pre-chosen probability of rejecting the null hypothesis, the p-value indicates the probability calculated after the study. If the p-value is less than the chosen significance level then null hypothesis is rejected.

Table 2.1
Decision rules for hypothesis tests.

	Real situation	
Decision	H_0 true	H_1 false
H_0 not rejected	No error $(1-\alpha)$	Error type II (β)
H_0 rejected	Error type I (α)	No error $(1-\beta)$

The decision rules of a hypothesis tests are summarized in table 2.1. The error type I, denotes the error of rejecting H_0 when it is true. Thus, the value α denotes the probability of an error of type I ($\alpha = P(reject H_0 | H_0 true)$), and it is also called significance level. The error type II denotes the error of not rejecting the null hypothesis when it is false. The value β denotes the probability of an error of type II. Finally, the power of a hypothesis test denotes the probability of rejecting H_0 when it is false. The power increases when the sample size increases, and it is called $1 - \beta$.

We can distinguish among two different types of hypothesis test, as figure 2.4 illustrates: one-tailed tests, where the region of rejection is on only one side of the sampling distribution, and two-tailed tests, where the region of rejection is on both sides of the sampling distribution.

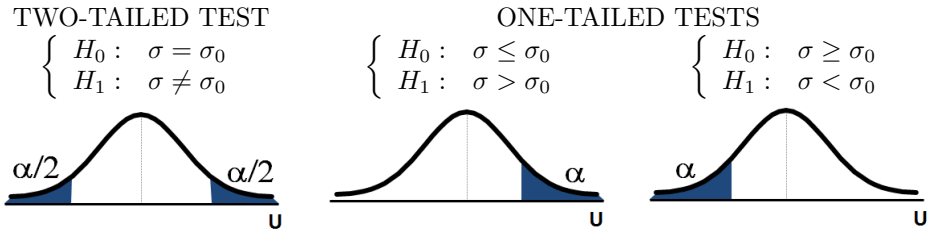


Figure 2.4: Types of hypothesis tests

2.4.3 Student's t-Test and ANOVA

A Student's t-test is a statistical hypothesis test in which the statistic follows a t-Student distribution if the null hypothesis is true. This test is generally used to determine wether there are significant differences between two sets of data. In order to apply the test the sample sets must be independent. A restriction of the t-test is that both sample population must follow a normal distribution. Note that the t-test does not prove anything, but can support a hypothesis. It is usually applied to small samples where you can not use more advanced techniques.

When there are more than two samples to compare among them, the results of several t-tests might become unreliable, and thus we need advanced techniques such as Analysis of Variance tests, ANOVA. ANOVA tests are used to analyze different kinds of variability in the data and then they use that information to construct a hypothesis test.

ANOVA analyzes data consisting of one quantitative response variable and one or more categorical explanatory variables, referred as factors. The number of factors depends on the situation to analyze. We will focus on the two simplest cases, using one or two factors with one response variable, which are called one-way and two-way ANOVA, respectively.

One-Way ANOVA

In general, a one-way ANOVA analysis considers the factor dividing the subjects into groups. Thus, the goal of the analysis is to compare the means of the subjects in

each group and the objective of this test is to select the best parameter settings. The results are considered reliable if the response variable residuals are normally distributed or approximately normally distributed, the samples are independent, the variances of populations are equal and the responses for a given group are independent and identically distributed normal random variables.

The hypothesis test associated to the one-way ANOVA considers as null hypothesis H_0 that the factor has no effect, and as alternative that it does. In terms of parameters, the ANOVA test can be written as follows:

$$\begin{cases} H_0: & \mu_1 = \mu_2 = \ldots = \mu_g \\ H_1: & \exists\, \mu_i \text{ s.t. } \mu_i \neq \mu_j \text{ for some } j = 1, \ldots, g \end{cases}$$

If the p-value is less than or equal to α, we can reject H_0, and then we conclude that at least some means of the group population are different. If we fail to reject H_0, that is, the p-value is greater than α, then we conclude that it is reasonable that all the group population means are the same.

Two-Way ANOVA

The two-way ANOVA test is used to detect the impact of the variability among two factors and allows us to get a more accurate description of how the response variable depends on the two factors, detecting any interaction across the different factors that might distort the analysis of results separately for each factor. In addition, the two-way ANOVA considers more sources of variability than each individual one-way ANOVA does, which leads to fewer errors of type II. The two-way ANOVA test can answer if the response variable depends on the first factor, or if it depends on the second one. Indeed, it also answers if the response variable depends on the first factor differently for different values of the second factor and viceversa. The assumptions for the two-way ANOVA test are the same as for the one-way ANOVA test, but also that the number of observations should be the same for all groups.

Suppose we have Factor 1 with N categories and Factor 2 with M categories. Then the total number of groups is NM. The first question to be answered is if the response variable depend on Factor 1 differently for different values of Factor 2, and viceversa. The interaction test stands the null hypothesis as there is no interaction, while the alternative one supposes there is. If the p-value is less than or equal to the significance level α, then we reject H_0 which means that is reasonable to think that there is interaction. Whereas, if p-value is greater than α then we fail to reject H_0, which means that it is reasonable to think that there is no interaction.

If we assume that it is reasonable that there is no interaction, the response variable still can depend on Factor 1 and/or Factor 2. Then we can look at the effects of both factors separately, so we test for the main effects. The main effect test behaves as the previously described test. That is, the null hypothesis stands that the response variable does not depend on the factor in question, while the alternative hypothesis is that it does. Thus, if the p-value is greater less than or equal to the significance level α, then we can reject H_0, and thus, then we can assume that the response variable depends on the selected factor. If the p-value is greater than α, we can not reject H_0 then it is reasonable to assume that the response variable does not depend on the selected factor.

If we can assume that there is interaction, it means that the response depends on the first factor differently for different values of the second factor and viceversa. And thus, there is no direct dependence of one factor or the other one. In this case, a 1-way ANOVA combining both factors as one should be applied to determine whether or not the response variable depends on the combined factors.

Multiple Comparison Methods

In case one factor is identified as influent, it is necessary to confirm what levels differ mutually. For that, several comparisons among the groups might be done. Since there is a significance level of 5%, there is a risk in each comparison of 5%, so that a large number of comparisons significantly increments this risk. Thus, some kind of correction should be done to have an overall risk below 5%. In particular, for pairwise comparisons Tukey correction [76] is the most appropriate for constructing the confidence intervals.

For each comparison of two groups, we interpret the corresponding Tukey simultaneous confidence interval as follows:

- If the interval contains only positive numbers, then we can conclude that the first of the two population means being compared is bigger than the second.
- If the interval contains only negative numbers, then we can conclude that the first of the two population means being compared is smaller than the second.
- If the interval contains both positive and negative numbers (in other words, if it contains zero), then we can't conclude that either of the two population means being compared is bigger than the other.

Of course, whenever we conclude that one population mean is bigger than another, the interval also gives us an idea of how much bigger.

Chapter 3

Descriptive Scores for Confidence Measure Performance Evaluation

In order to use optical flow in a confident decision support system, a mechanism to detect sequence pixels that have high error in their computations is of prime importance. In this context, Confidence Measures (CM) should be an indicator of the accuracy of the output of an optical flow algorithm. Thus, evaluating the quality of a confidence measure, as well as analyzing the origin of unpredictable points, are issues as important as the definition of a confidence measure itself.

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. 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 [49].

This chapter is devoted to the analysis of OF error sources and the capabilities of CM for bounding such OF errors. On one hand, we describe the different error sources of OF techniques and the theoretical capabilities of existing CM types for bounding the different sources of OF errors. On the other hand, we contribute to the quantification of the decreasing pattern of CM-OF error point clouds in two aspects. First, we present 3 descriptive plots reflecting the decreasing profile of the CM-OF error point clouds in terms of the distribution function of points that CM can not bound OF error. Each plot increasingly approaches different aspects that contribute to a better qualitative description of the CM-OF decreasing profile and leads to the definition of our definitive Sparse Density Plots, SDP. Second, we provide a set of scores actually quantifying such decreasing profile from the analysis of descriptive plots. The higher descriptive capabilities of our plots in comparison to the existing Sparsification Plot, SP, have been exposed in two representative sequences extracted from the benchmark Middlebury and Sintel databases. The usefulness of the descriptive scores for using SDP in a decision support system is shown by applying the whole

description framework to the full Sintel database.

3.1 Theoretical Capabilities for Error Bounding

The final purpose of a confidence measure should be the detection of numerical errors and provide an upper bound for the final error. Therefore, before introducing the 3 plots describing the capabilities of CM for OF error bounding, let us analyze the different sources of error that an OF algorithm has and that CM should be able to detect.

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 [77]. In the second case, the flow vector should follow a given probabilistic distribution or be generated by a finite number of basic functions [77]. 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. 3.1 (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 [35]), 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 [77], 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 3.1(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 modelling 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.3.1(c)). In mathematical numerical analysis this is called error propagation or numerical stability [78]. 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 [78]. 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. 3.1(c)).

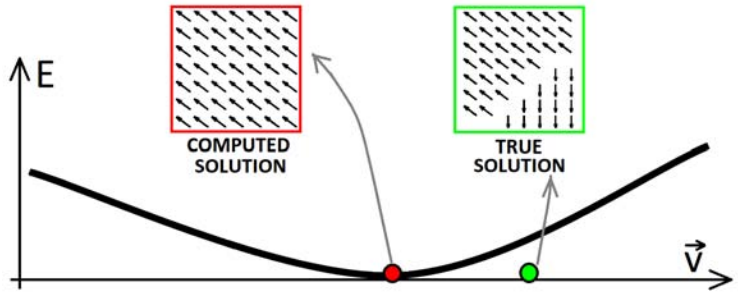
Taking into account the main sources of OF error described above, current CM have different theoretical capabilities as evaluation tools. We recall that (see Chapter 2) confidence measures can be formulated from either an analytic or a probabilistic point of view. Analytic approaches either use the energy [49,50] or the image structure (gradient magnitude [31], structure tensor [51]) as indicators of confidence. Whereas, probabilistic approaches define confidence in terms of probabilistic distributions of either flow fields itself [52] or its variability with respect perturbations in the model [53]. Considering their formulation, described in Chapter 2, we have the following categorization (summarized in Table 3.1) for error bounding.

Analytic Formulations. Energy-based measures [49] evaluate OF in terms of the pixel-wise value of the integrand defining the variational that OF solves:

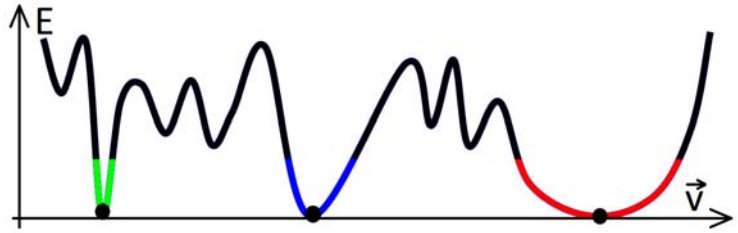
$$C_e = \frac{1}{D(u, v, \nabla I) + \alpha S(\nabla u, \nabla v, \nabla I) + \epsilon^2} \tag{3.1}$$

where ϵ prevents dividing by zero. A main advantage of C_e is that it can be computed for any variational scheme. A main concern is that C_e only measures that (u, v) minimizes equation (2.6) and, thus, fulfills the assumptions made in the model. However, this does not guarantee that (u, v) corresponds to the true flow field, since defining the most appropriate optical flow constraints for a given application is still an open problem.

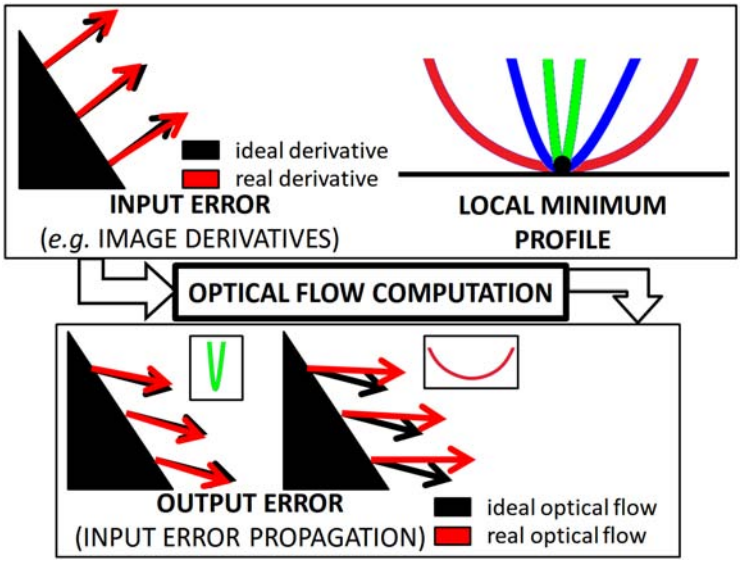
The other big group of confidence measures having analytic formulation are those measures defined by means of the structure tensor of the image, and thus, they use information about the local structure of the image. There are several measures derived from the structure tensor: determinant [31], trace [75], lowest eigenvalue [51] among others. These measures only detect errors produced due to the image, that is, textureless regions, noise, etc. However, they do not consider the errors produced by during the computations. An improved measure that uses the structure tensor,



(a) Model assumptions.



(b) Multiple global minima.



(c) Numerical stability of local minima.

Figure 3.1: Three main sources of error in optical flow algorithms: model assumptions (a), multiple global minima (b) and numerical stability of local minima, (c).

considers the condition number of the structure tensor matrix [79]:

$$C_k = \frac{\lambda_{min}}{\lambda_{max}} \tag{3.2}$$

for λ_{min} and λ_{max} the minimum and maximum eigenvalues of the structure tensor at a given pixel location. This measure, not only assesses the capabilities of the image to compute the flow field, but also assesses the numerical stability of the computation for Lucas-Kanade based schemes [33, 38].

Probabilistic Formulations. In many applications, flow fields follow similar local motion patterns. If such motion patterns are learned a priori, then a classifier can be used to define a confidence measure. The measure introduced in [52], which we note as C_s, derives natural motion statistics from sample data and carries out a hypothesis test to obtain confidence values for the computed flow. The method depends only on the resulting flow field and on the prior knowledge learned from a database. The confidence measure 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.

A main limitation of this measure is that unusual motion patterns are not easy to learn and might require a huge database of different flow patterns to train the model. This limits its use for sequences with flow fields that are erratic or unpredictable. In addition, it only assesses if the flow field is coherent, but not if the flow field corresponds to the sequence motion.

The measure introduced in [53] aims at assessing the uncertainty of the optical flow method with respect to the model constraints. That is, they compute the variability of the computed flow field using bootstrap introducing numerical perturbations. If the variability is high, the flow field is not reliable, whereas for low variability, the computation is reliable. In order to have a decreasing dependency with the accuracy is rewritten follows:

$$C_b = \frac{1}{\psi_{bootg} + \epsilon^2}, \quad \psi_{bootg} = \sqrt{\sigma_u^2 + \sigma_v^2} \tag{3.3}$$

for ψ_{bootg} the confidence measure defined in [53], eq. (15). Like C_e, this one also assesses the consistency of the model assumptions, but also assesses the errors produced by numerical stability of the method. However, C_b requires to be redefined for each optical flow technique and it is computationally costly.

Finally, the measure presented by Aodha et al. [54] estimates if a flow algorithm is likely to fail in a specific region by means of supervised learning. This confidence measure can be computed fusing all previous measures and, thus, it can relate to all three error sources.

3.2 Descriptive Plots

Scatter plots showing CM against OF errors are a good tool to assess the relation between both quantities, as illustrated in fig.1.4. A perfect CM should produce decreasing profiles, like the one shown in fig.1.4(a). In fig1.4(a) points over the red dashed line to points which error is not bounded by the confidence measure. Given that these points could introduce a significant error in a decision support system using OF, evaluation of the confidence measure should detect the scope of such unpredictable points.

			Error Type		
			Model Assumptions	Multiple Minima	Numerical Stability
Measure Formulation	Analytic	gradient-based [31]	✓	×	✓
		image local structure [51]	✓	×	✓
		energy-based [49]	×	✓	×
	Probabilistic	bootstrap [53]	✓	×	✓
		p-val [52]	×	×	×
		random forest [54]	✓	✓	✓

Table 3.1
Categorization of confidence measures according to error types.

We have been talking about the importance of a confidence measure, but, what exactly should we expect from them? We expect a decreasing dependency between the measure and the accuracy. That is, for high values of the measure, we expect high accuracy, whereas for low values we expect lower accuracy. However, looking at the scatter plot between the accuracy and CM we observe that they follow a probabilistic distribution rather than a pure analytic function. Therefore, we state the decreasing dependency between confidence measure CM and error EE using the following inequalities to define our Condition of the Quality Threshold (CQT):

Definition 1 Condition of the Quality Threshold (CQT). *We say that a CM is good for OF error bounding if for any CM value we have a threshold for EE, that is:*

$$\forall cm, \exists ee \text{ such that } CM > cm \Rightarrow EE < ee$$

That is, for a given probability, α, the ideal confidence measure should be able to guarantee that for a threshold cm on CM, the error, EE, is bounded. We note that under CQT, the values of confidence measures would determine the accuracy of the flow field in the absence of ground-truth.

3.2.1 Error Predictive Plots

Bearing the above requirements in mind, we propose the following confidence framework on the grounds of numerical stability analysis:

The CQT is fulfilled only if the scatter plot between a confidence measure CM and an error EE show a perfect decreasing pattern. Such pattern is difficult to measure using mathematical analysis tools because they are unable to properly handle point distributions. The best way to explore point distribution is by means of probability density functions. In probabilistic terms CQT can be stated as the following conditional probability:

$$P(EE \geq ee | CM \geq cm) < \alpha \tag{3.4}$$

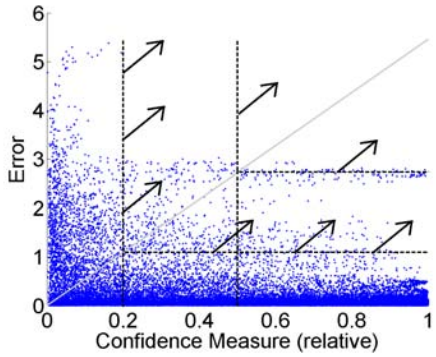


Figure 3.2: Probability density function of CM capabilities for OF error bounding

for $\alpha \leq 1$ the probability of having an error above ee provided that CM is above cm. The conditional probability can be computed by scanning the scatter plots given by CM-EE as follows: taking into account that the condition $CM \geq cm$ corresponds to a vertical line and $EE \geq ee$ to an horizontal one, the conditional probability is given by the fraction of points lying on the superior quadrant defined by the former lines.

The plot in fig.3.2 illustrates the computation of (3.4) by fully scanning the CM-EE space. Arrows indicate the points that are considered for the computation of conditional probabilities. The conditional probability (3.4) is a bi-dimensional graph, not easy to interpret. In order to get a simpler representation able to assess the capability of the measure for predicting the error, it suffices to consider the values for the diagonal of the support domain of CM-EE plots. Such line is given by the following equation:

$$ee = cm\frac{max(ee)}{max(cm)} \tag{3.5}$$

for $max(ee)$, $max(cm)$ the expected maximum values of CM, EE computed for a training set. We note that the conditional probabilities:

$$P(EE \geq cm\frac{max(ee)}{max(cm)}|CM \geq cm))$$

increase in case we have points with a high EE and CM. Therefore, by scanning all diagonal values, we ensure that unusual non-decreasing patterns (as the one shown in the left scatter point cloud in fig.3.3) are detected. We define our Error Prediction Plots, EPP, as the plot given by:

$$EPP := cm \mapsto (cm, P(EE \geq cm\frac{max(ee)}{max(cm)}|CM \geq cm))) \tag{3.6}$$

Figure 3.3 shows scatter plots and their corresponding EPP plots. Unlike the sparsification plot shown in 2.3, we observe that EPP is worse for the non-decreasing case. In this manner, as illustrated in 3.3, EPP partially overcomes SP limitations

and, by definition, they are better designed to detect artifacts in CM-OF error scatter plots. However a main concern is the sampling of the 2D distribution space which is partially scanned and implies an increasing OF error as CM values are swept. In order to skip scanning the full 2D space, we propose summarizing CM-OF error scatter plots in a different manner by reformulating CQT condition given in Definition 3.2 as follows:

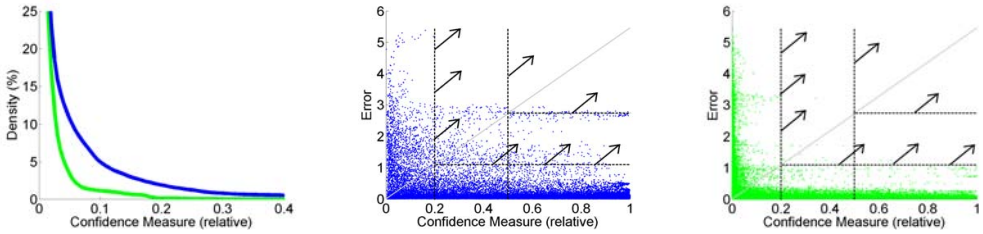


Figure 3.3: Error Prediction plots. First column shows the EPP plot for two different scatter plots. Second column shows the scatter plot of a confidence measure versus the error. On the left a poor measure, on the right a good one.

3.2.2 RAUC plots

The condition CQT states that CM should give an upper bound for EE everywhere. That is, $\forall cm$, EE values should be bounded for all CM values above cm. In probabilistic terms, this implies that the following conditional probability is zero:

$$\forall cm, \exists ee \text{ such that } P(EE > ee | CM > cm) = 0 \tag{3.7}$$

Figure 3.4 (a) illustrates ee error bounds for different cm values in the case of a perfect relationship between CM and EE. Vertical lines correspond to cm values and horizontal lines the best ee bound for such cm values.

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

$$\exists cm \text{ s.t. } \forall eee, P(EE > ee | CM > cm) > 0 \tag{3.8}$$

We define the risk of a confidence measure as the proportion of points, ρ, which bound can not be determined by CM values:

Definition 2 Risk. *We define the risk of a confidence measure as*

$$\rho := P(EE \geq ee | CM \geq cm)$$

$$ee_{opt}(cm) := min\,(ee \ such \ that \ \rho(ee) \equiv 0)$$

The scatter plot in fig. 3.6 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 at $CM = cm$ ($\rho_1 > \rho_2 > \rho_3$). For each EE its risk is given by the percentage of points on the upper right square defined

by the two lines. Given that EE can take any possible value and this holds for all $CM > cm$, 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. 3.4(b), where, for each CM value, equation (3.7) 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 for error bound setting, since the only requirement for error bounding is that curves should be monotonically decreasing. Figure 3.4(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 labelled 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 $ee^1_{opt} < ee^2_{opt}$. 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, so that points above it represent the risk associated to c_1.

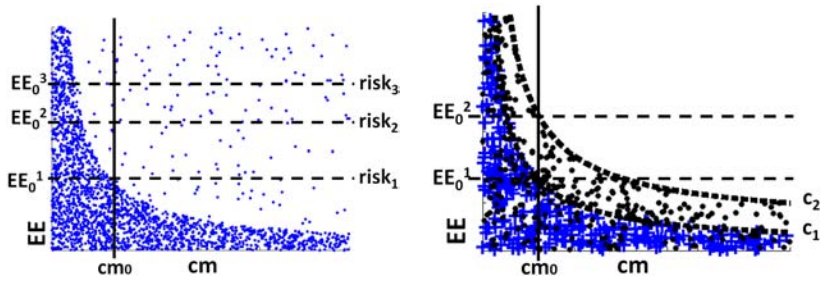


Figure 3.4: Concepts involved in the quality of a confidence measure: risk in error bound prediction, (a), and optimal error bound for a given risk, (b).

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 should provide the lowest possible bound ee. 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 ($\rho(ee)$) and ee_{opt} 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. The convex curve enclosing the maximum number of pixels gives the

lower bound for *ee*. Its risk is a measure of the maximum risk for the best *ee* bound. The intermediate curves are computed iteratively removing a percentage of these points as follows.

As previously described, confidence curves should be monotonically decreasing and minimize the risk of false error bounds. To achieve this, several approaches are possible. As the number of points in an evaluation is finite, one (optimal) option would be to combinatorially compute all possible monotonically decreasing curves given a dataset. The number of points is larger than a million, so this method takes an infeasible amount of time to find the optimal curves.

To approximate the approach above with as little loss of optimality as possible, we first create a two-dimensional histogram H which serves as an approximate Parzen density [80] estimate of the dataset. To obtain curves which minimally increase the risk each step similarly to the sparsification plot, we remove a fraction of bins by setting them to zero. We then search for a monotonically decreasing function in this set of remaining points with all point being below the curve. The choice of bins is based on the principle of minimizing the area under the resulting curve while also minimizing the risk increase caused by the points which are removed (and therefore might be located above the curve). We therefore use a prioritized recursive region growing approach with one seed at the location of both maximal error and confidence (top right corner). The recursion stack is ordered by the number of points which fell into the bin of the histogram. At lower priority, bins with the same value are ordered by the distance from the seed point. At lower priority, bins with the same value are ordered by the L^1-distance from the seed point. In case all bins have the same value, this creates a linear (and therefore monotonically decreasing) confidence curve.

We iterate through the recursion by removing bins until the sum of the removed bins reaches a fraction of the sum of all bins. Each curve will thereby yield a good approximation of the optimal confidence curve given an acceptable, application-dependent risk defined by the user. As described previously on this section, the area under the curve can then be used to define the quality of the confidence measure.

Figure 3.5 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*). It is ideal because it is able to produce a convex curve achieving zero risk. Consequently, its RAUC first point starts with the lowest 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.

Although the concept of risk allows faithfully summarizing CM-OF error scattered plots, the proposed RAUC is not invariant under transformations of the confidence measure, which do not alter its error bounding capabilities. In order to obtain plots

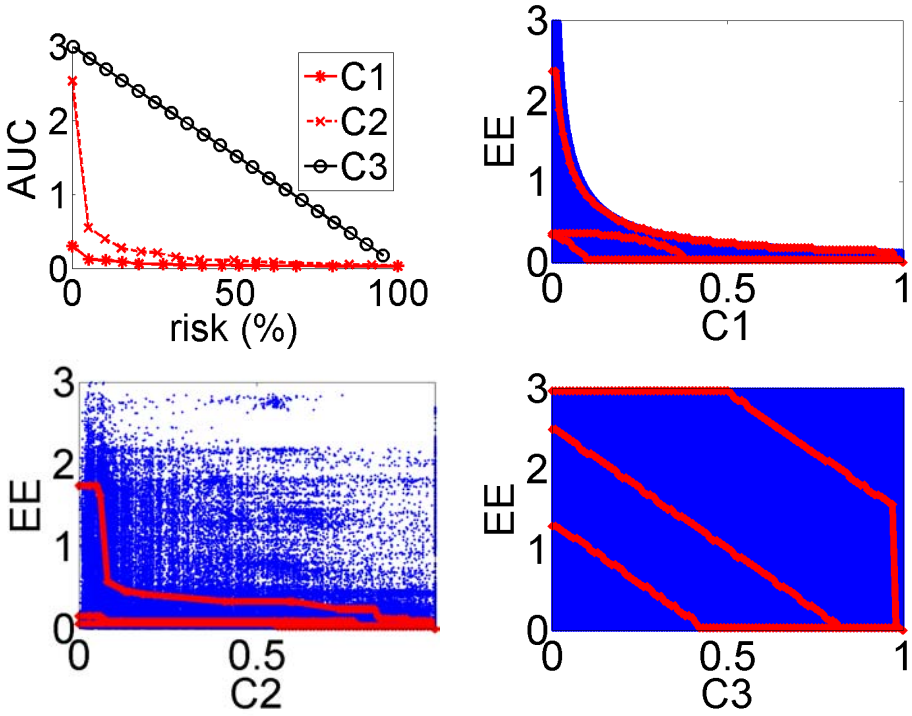


Figure 3.5: Synthetic example. First column, the RAUC for the three different cases, ideal case ($C1$). Second column, expected case ($C2$) and worst case ($C3$).

invariant under CM scalings, we will use the sparsified predictive value over CM percentiles to define our Sparse-Density Plots (SDP) as follows:

3.2.3 Sparse-Density Plots

Considering CM as a diagnostic score, the risk associated to CQT can be assessed using the positive, PV_+, and negative, PV_-, Predictive Values [81]. These scores are widely used in medical diagnostic tests and are given by the following distribution functions, $PV_+(cm)$, $PV_-(cm)$, computed considering CM as a random variable:

$$\begin{aligned} PV_+ = PV_+(cm) &:= P(EE \leq EE_{max} | CM > cm) \\ PV_- = PV_-(cm) &:= P(EE > EE_{max} | CM \leq cm) \end{aligned} \tag{3.9}$$

for EE_{max} the maximum error allowed in our application. In the context of decision support systems, it is more descriptive to consider the probability of points whose bound can not be determined by CM:

$$1 - PV_+(cm) = P(EE > EE_{max} | CM > cm) \tag{3.10}$$

We note that such points act as outliers that should either be discarded or processed separately. Thus, we will call $1 - PV_+(cm)$ risk and will note it by $\rho = \rho(cm)$. The

CM-EE scatter plot in figure 3.6 illustrates the concept of risk. The vertical dashed line at CM represents the threshold for CM and the horizontal dashed line the bound on EE given by EE_{max}. For each CM value, its risk is given by the percentage of false positive points over the addition of false positive and true positive points.

The risk plots, $(cm, \rho(cm))$ provide information about CM capabilities for error bounding. In order that we get plots invariant under monotonically increasing transformations (which do not alter CM bounding capabilities), $(cm, \rho(cm))$ plots are sampled using the percentiles of CM distribution. If we note such percentiles by $prct_{CM}$, then we define our Sparse-Density Plots (SDP) as:

$$SDP : prct_{CM} \mapsto (prct_{CM}, \rho(prct_{CM})) \tag{3.11}$$

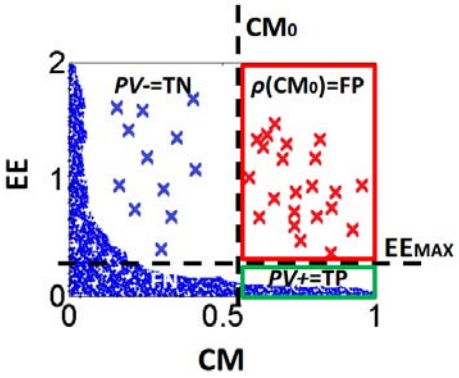


Figure 3.6: Scatter plot between CM and EE. Dashed vertical line shows the bound of CM (CM_0) and horizontal dashed line shows the maximum EE allowed (EE_{max}).

Figure 3.7 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 fig3.7(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.3.7(b) and (c) come from the most usual CM behaviors. In the first case (fig.3.7(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.3.7(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.3.7(d) and (e) show the worse cases, in the sense that CM is not related to OF error. The constant profile of fig.3.7(d) indicates that the $CM - EE$ distribution is uniform and, thus, EE can take any value regardless of CM. The case shown in fig.3.7(e) is even worse. It has a behavior opposite to the expected one as large CM values have an unbounded error.

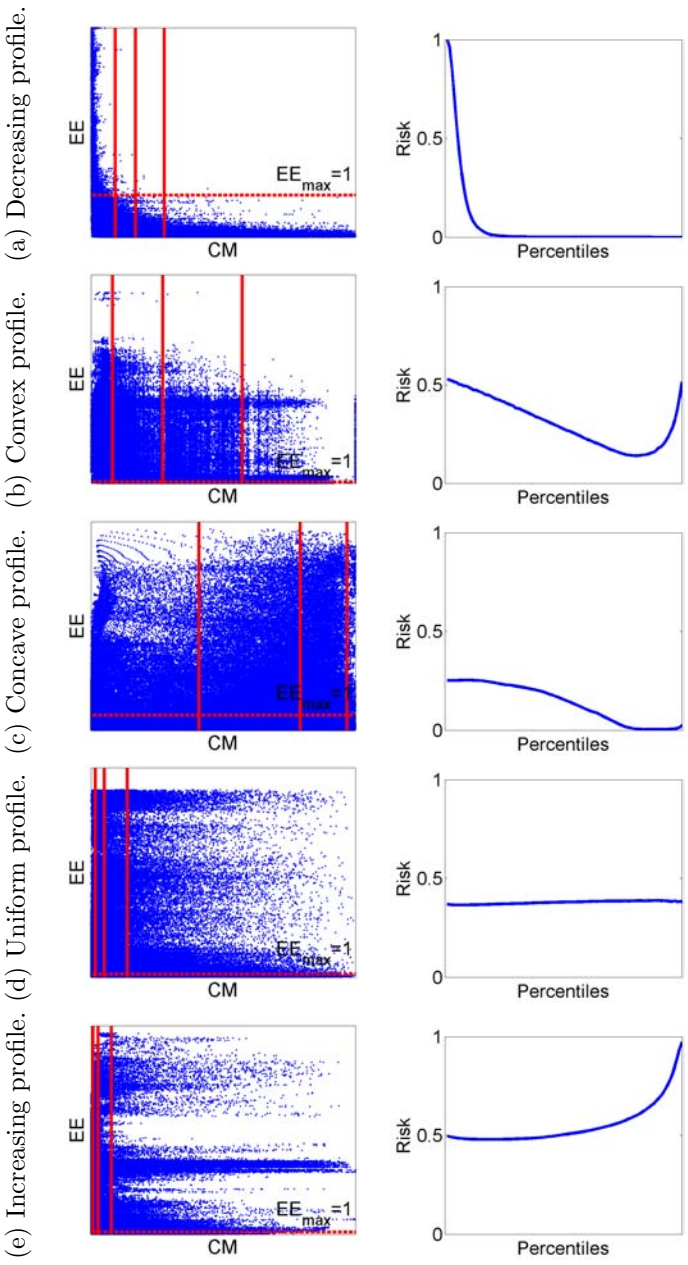


Figure 3.7: Representative examples of different SDP, ranged from best to worst capabilities for error bounding. Left column shows to the scatter plots (CM vs EE). Vertical red lines correspond to the percentiles $0.25, 0.5, 0.75$ and horizontal red line indicates the $EE_{max} = 1$. Right column shows the respective SDP.

3.3 Descriptive Scores

To properly bound OF error, the profile of descriptive plots should be monotonically decreasing and in fact, non monotonically decreasing profiles arise in the presence of risk points. Figure 3.8 shows the different profiles sorted in decreasing capabilities for error bounding. Plots in first column show the best profile and plots in last column the worst one. The first row shows CM-EE scatter plots for a single frame with the percentiles $\{0.25, 0.5, 0.75\}$ and $EE_{max} = 1$ in solid and dashed red lines respectively, and the second row shows the corresponding descriptive plot. We label the descriptive plot bounding capabilities with $3, 2, 1, 0, -1$, being 3 best and -1 worst, according to the three following conditions: proportion of decreasing points, range of decrease and CM percentile decreasing points. The mentioned conditions and a Descriptive Plot (DP) classification are illustrated in fig. 3.8 and can be formulated as follows:

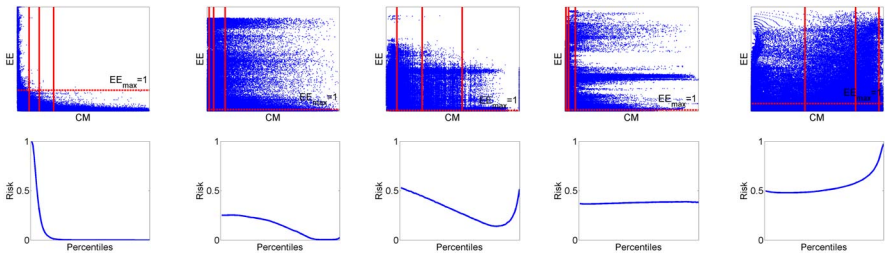
	Profile				
	(a) Decreasing	(b) Concave	(c) Convex	(d) Uniform	(e) Increasing
DP profile Scatter CM-EE					
DP_{Lab}	3	2	1	0	-1
$Cond_1$	1	1	1	1	0
$Cond_2$	1	1	1	0	-
$Cond_3$	3	2	1	-	-

Figure 3.8: Illustrative examples of different descriptive plots, ranged from best (left) to worst (right) capabilities for error bounding: CM-EE scatter plots in 1st row, descriptive plot curve in 2nd row and its categorization in the last row.

- **Proportion of descriptive plots decreasing points.** For descriptive plots increasing profiles, large CM values have unbounded error. This is the worst situation, opposite to the expected behavior (see figure 3.8(e)), and, thus, descriptive plots should be assigned label -1. Descriptive plots increasing profiles are detected by means of the sign of the descriptive plot first derivative, which should be negative in a large enough number of $prct_{CM}$ points. Let $X_{DP}(prct_{CM})$ be the sign function of the descriptive plot first derivative:

$$X_{DP}(prct_{CM}) = \begin{cases} 1 & \text{if } DP'(prct_{CM}) \leq 0 \\ 0 & \text{otherwise} \end{cases} \tag{3.12}$$

which is 1 at points where the profile is decreasing and 0 otherwise. We consider that the plot has a non-increasing profile if the proportion of points having a positive X_{DP}:

$$q_{DP} = \frac{\sum_{prct_{CM}} X_{DP}(prct_{CM})}{|prct_{CM}|}$$

is above a given tolerance q_1:

$$Cond_1 = \begin{cases} 1 & \text{if} \quad q_{X_{DP}} \geq q_1 \\ 0 & \text{if} \quad q_{X_{DP}} < q_1 \end{cases} \tag{3.13}$$

Note that $1 - q_1$ is the maximum percentage of pixels which error cannot be bounded by CM and it is application dependant. In case $Cond_1 = 0$, the profile is considered increasing and it is assigned the worst label equal to -1. Figure 3.9 (a) illustrates the computation of the proportion $q_{X_{DP}}$ with DP decreasing intervals highlighted in red.

- **Descriptive plot range of decrease.** Curves with uniform profiles (like the one shown in figure 3.8(d)) indicate that $CM - EE$ distribution is uniform and thus, EE can take any value regardless of CM. Such cases are assigned a label 0 and are detected evaluating if the descriptive plot decrease range:

$$Rng_{DP} := DP_{max} - DP_{min}$$

is above a given percentage, q_2, of the plot maximum value,

$$DP_{max} := max_{prct_{CM}}(\rho(prct_{CM}))$$

and thus, the second condition is defined as:

$$Cond_2 = \begin{cases} 1 & \text{if} \quad Rng_{DP} \geq q_2 DP_{max} \\ 0 & \text{if} \quad Rng_{DP} < q_2 DP_{max} \end{cases} \tag{3.14}$$

Like q_1 in the first condition, the decreasing proportion q_2 depends on the decision support system requirements. Figure 3.9 (b) illustrates the computation of the range condition (3.14).

- **Location of descriptive plot decreasing points.** We observe that a confidence measure is able to completely bound OF error if a descriptive plot has a strictly decreasing profile and reaches the zero value for some $prct_{CM}$, like the profile shown in figure 3.8(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 figures 3.8(b) and (c) come from the most usual CM behaviors. In the first case (figure 3.8(b)), there is a group of pixels with unbounded errors in the first CM percentiles but for higher percentiles, the error is completely under control. In the second case (figure 3.8(c)), 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 descriptive plot graphics.

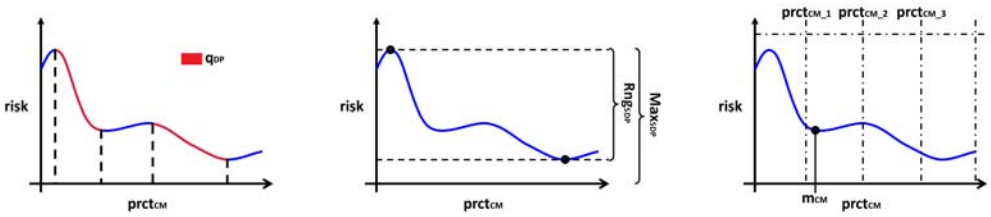
Therefore, the descriptive plot label is 3, 2 or 1 according to the $prct_{CM}$ value which the plot begins to be increasing. Let m_{CM} be such point defined as:

$$m_{CM} = \min\{prct_{CM} \| DP' > 0\}$$

and consider a partition of the interval $[0, 1]$ given by $I_3 = [0, q_{33})$, $I_2 = [q_{33}, q_{32})$ and $I_1 = [q_{32}, 1]$ (application dependant as before). Figure 3.9 (c) illustrates the computation of this final label according to a partition given by the dashed vertical lines.

The final plot label is given by:

$$DP_{Lab} = \begin{cases} i & m_{CM} \in I_i \ , \ Cond_2 = Cond_1 = 1 \\ 0 & Cond_2 = 0 \ , \ Cond_1 = 1 \\ -1 & \text{otherwise} \end{cases} \tag{3.15}$$



(a) $Cond_1$ (b) $Cond_2$ (c) $Cond_3$

Figure 3.9: Graphical representation of the conditions $Cond_1$, $Cond_2$ and $Cond_3$.

3.4 Experimental Settings

The goal of these experiments is to validate the analysis of confidence measure capabilities and also the descriptive plots and scores presented in this chapter as tools for assessing the quality of a given CM for bounding OF error. The following experiments have been carried out:

1. **Analysis of confidence measures and descriptive plots capabilities.** This experiment has two purposes, analyze the confidence measures capabilities depending on the optical flow error source, and to analyze which descriptive plot can assess better the scatter plot CM-EE.

   On one hand, with this experiment we pretend to better understand a confidence measure behavior and its weak and strong points for bounding OF error. In particular we will address the local conditions (both in appearance and motion) that a sequence should fulfill in order that a CM succeeds in bounding the error of a particular OF method. In order to explore CM bounding capabilities we locally analyze the behavior of confidence measures for a selected sample of sequence patches. These patches cover the main appearance and motion features that are prone to introduce an error in OF and CM expected behaviors. In this context we have selected patches violating:

- **Data-term OF constrain assumptions.** On the one hand, the data term requires that there is enough information in the image intensity to compute the apparent 2D motion. On the other hand, large displacements are against the first order Taylor approximation given by the OF equation. Therefore, we have selected patches with straight edges and textureless regions for their intensity appearance as well as, patches of a large displacement.
- **Smoothness-term regularity assumptions.** Independent motions might interfere with the regularity assumptions of the smoothness-term. Thus we have selected regions with motion discontinuities.

On the other hand, we have explored the capabilities of the three plots (EPP, RAUC and SDP) presented in section 3.2 as tools for assessing CM quality in terms of OF error bounding power in comparison to Sparsification Plots (SP). This exploratory analysis will select the descriptive plot having the highest capability for assessing CM quality. The selected plot will be used for further experiments.

In order to carry out the experiments we have chosen two representative examples of confidence measures of each formulation category. The structured based measure [79], and the energy based [49] for analytic formulations and statistical [52] and bootstrap [53] for probabilistic approaches. They are denoted as C_k, C_e, C_s and C_b, respectively (see section 3.1 for more details). Optical flow has been computed using the Combined Local-Global (CLG) scheme [38] as implemented in [82] [1]. The error score is the End-Point Error (EE) [48], which measures the difference between computed flow field and ground truth. Results have been extracted from two benchmark databases (Middelbury [48] and Sintel [68]) (for more information see Chapter 2). For the first part of the experiment, we have chosen the sequences RubberWhale and Urban from Middlebury dataset. For the second part of the experiment we have chosen the sequences Hydrangea and Grove3 from Middlebury dataset, and cave2 and sleeping1 sequences from Sintal database.

2. **Assessment of the descriptive scores.** The second experiment is an explorative analysis of predictable sequences by means of the descriptive scores computed for the descriptive plot selected in the first experiment. To assess the descriptive scores, we have computed the plot label for each sequence frame and taken the median as summarizing score. The parameters used to compute the labels are $q_1 = 0.75$, $q_2 = 0.8$ and $q_{33} = 0.25$, $q_{32} = 0.75$.

   In this case, we have used all Sintel sequences and extended the OF methods in order to cover the following representative and state of the art optical flow methods[2]. The classic formulations are: Combined Local-Global (CLG) [38], which uses a Lucas-Kanade data term [33] with an L^2 norm smoothness term and Horn-Schunck (HS) [35], which uses OF brightness constancy equation with an L^2 norm smoothness term. Meanwhile the state of the art are the

[1] Available at http://people.csail.mit.edu/celiu/OpticalFlow/
[2] Using the free source code from [82], [40] and [67].

Classic-NL (NL) [40], which is a total variation method that uses the L^1 norm to combine OF brightness constancy assumption with the smoothness term and Correlation method ($Corr$) [67], which uses an L^2 data term based on the correlation transform of the images with an L^1 regularity term based on bilateral filtering (see Chapter 2 for more information). Concerning confidence measures, we have considered the same four as before. Therefore, we have $3 \times 4 = 12$ possible OF-CM pairs.

For all experiments, all CM are assumed to take values in the interval $[0, 1]$, 0 meaning low confidence and 1 high confidence. By definition, this is the range of C_k and C_s, but, C_e and C_b have to be normalized. This normalization can be a global one given by CM definition (such as C_e) or an empirical one computed over a sample (such as C_b). In the case of C_e we have changed its formulation to $C_e := \frac{1}{C_e+1}$ to ensure that C_e values are in the range $[0, 1]$. In the case of C_b, which always has positive values, we empirically normalized it by using the training sample, that is, $C_b := \frac{C_b}{max(C_b)}$, where $max(C_b)$ is the maximum value of the confidence measure for all points of the set.

3.5 Results

3.5.1 Analysis of confidence measures and descriptive plots capabilities

Figures 3.10 and 3.11 show our analysis for RubberWhale and Urban3 sequences. Each figure shows a sequence frame, 4 representative patches with computed (yellow arrows) and ground truth (green arrows) flows and CM-EE scatter plots for each measure. Each patch is of size 7×7 and it is centered at the respective illustrative point shown in the sequence frame and scatter plots.

The patches selected for fig.3.10 contain straight edges with independent motions (patches 1 and 4), a textureless region with uniform motion (patch 2), and a textureless region with a slightly irregular motion (patch 3). Figure 3.11 shows a sloped border with a large displacement of an object moving over a static background (patch 1), textureless regions with uniform motions (patches 2 a 4) and a slightly textured region with uniform motion (patch 3).

At straight edges (patches 1, 4 in fig.3.10) and textureless regions (patch 2, 3 in fig.3.10 and 2,4 in fig.3.11) CLG can not solve the data term. The lowest eigenvalue of the structure tensor matrix is close to zero and this introduces large numerical instability. We would like to note that in such cases EE can take any value, ranging within 0.4 and 10 in our sequences. Numerical instability of the data term is properly detected by low C_k values. The data term is numerically well-conditioned in the case of sloped borders (patch 1 in fig.3.11) and textured regions (patches 3 in fig.3.10 and 3.11). However, stable numerics do not guarantee accurate OF, given that OF model assumptions are also decisive for its accuracy. This is the case of patch 1 in fig.3.11, which presents a high error due to the high displacement magnitude and, thus, C_k can not properly bound EE. The bounding capabilities of C_b are more related to model assumptions and, thus, it properly bounds EE at patches presenting independent

moving objects (like patches 1, 4 in fig.3.10 and patch 1 in fig.3.11). However, its capabilities for error bounding decrease for patches with uniform motion, given that C_b is always high, but OF might present a large error for textureless patches (patch 2 in fig.3.10). The measure based on energy minimization, C_e, is also associated to model assumptions, given that at pixels which model regularity is not met, the functional can not be properly minimized. This is the case of patches with independent motions, like patches 1, 3, 4 in fig.3.10 and patch 1 in fig.3.11. Like C_b, C_e fails in the case of textureless regions with uniform motion shown in patch 2 in fig.3.10 and patch 3 in fig.3.11. Finally, the weakest measure for error bounding purposes is C_s, which scatters present the most uniform distribution of all. Such uniform distribution of EE across C_s values indicates that there is not a clear relation between the measure and the error. In fact, it only succeeds in bounding EE for patch 3 in fig.3.10 and patch 1 in fig.3.11 that are the ones having a flow field not regular around the central point. For the remaining patches, OF is regular enough although this does not necessarily imply it is accurate.

The second part of this experiment analyzes the capabilities of the different descriptive plots to detect when a confidence measure can not bound the OF error. Figures 3.12, 3.13, 3.15 and 3.14 show some example over benchmark sequences (Hydrangea, Grove1, Sleeping1 and Cave2 respectively). For each sequence, we show the four $CM-EE$ scatter plots with the different space partitions (depicted in black) that allow the computation of the descriptive plot. We also show the different descriptive plots for all CM. Note that in general, when a confidence measure shows a decreasing profile, the descriptive plot curve represents a good profile too. However when the scatter plot does not show a decreasing profile, then different descriptive plots show different rankings for different confidence measures. For instance, in fig.3.15, C_b scatters show a decreasing profile, and the different descriptive plots reflect such tendency. On the other hand, C_k does not always have a decreasing profile and such profile is not properly reflected by the RAUC curve.

Also note that the bounding artifacts detected in C_e profile in fig.3.14 are not properly reflected by SP plots. On the contrary, C_b decreasing profiles do not always produce a best decreasing SP (fig.3.14). This is due to only considering 1-dimensional statistics over EE and not over the bimodal distribution given by (CM, EE). By considering bimodal statistics, EPP provides a better ordering of CM quality for error bounding. In particular, it detects the groups of unpredictable pixels introducing horizontal scatters in CM-EE plots, like the ones present in C_b and C_e scatter plots in fig.3.14.

Figure 3.13 shows that the C_k scatter plot has a good EPP and RAUC profiles, similar to C_e and C_s, however, when you look at the scatter plots, the C_k performance is much worse than C_e and C_k. The bad profile of the scatter plot is better reflected by SDP and SP. Another example is shown in fig.3.12, where SDP reflects better capabilities to assess the performance of C_e than SP.

3.5.2 Assessment of the descriptive scores

A first analysis of the predictability of the sequences is summarized in table 3.2. For each sequence and each pair OF-CM, a label ranging from -1 to 3 is assigned to the

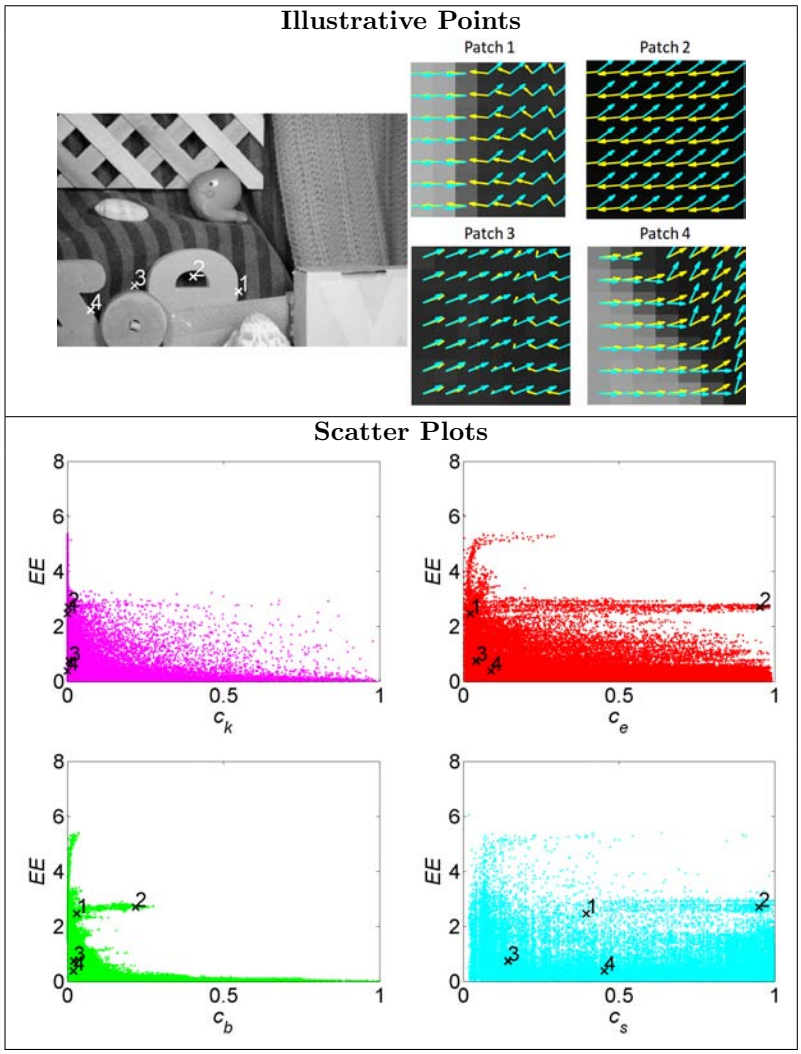


Figure 3.10: RubberWhale sequence.

profiles of the trained upper bound. The labels are assigned following the opposite order of the plots shown in figure 3.7 from the worst profile (-1) to the best one (3). To make the table more readable, we have assigned a different color to each label.

When we categorize the capabilities of the CM through the SDP profile, we can find three types of CM performances on sequences: *too good performance*, *too bad performance* (labelled by -1) and *predictable* ones.

Too good performance: 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

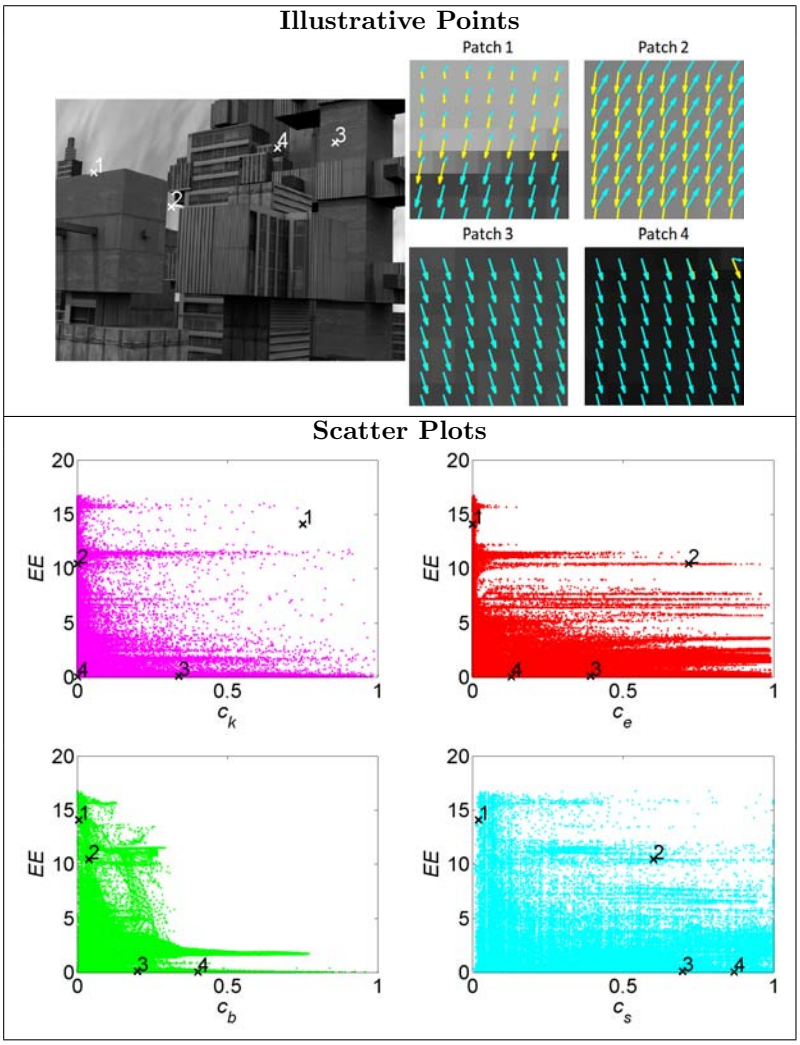


Figure 3.11: Urban3 sequence.

a very low upper bound 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. Such sequences are also labelled with 3. Figure 3.16(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 performance: Different OF methods require specific assumptions in order to properly compute the flow field. In case sequences do not fulfil such require-

	Ck				Cb				Ce				Cs			
	CLG	HS	NL	Corr	CLG	HS	NL	Corr	CLG	HS	NL	Corr	CLG	HS	NL	Corr
	$\mathcal{SDP}$ Predictive Value															
$alley_1$	0	0	0	-1	3	-1	3	0	3	3	3	3	-1	1	-1	1
$alley_2$	0	0	0	0	-1	3	-1	-1	-1	3	-1	-1	3	-1	-1	3
$ambush_4$	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
$ambush_5$	0	0	×	×	×	×	×	×	×	×	×	0	×	×	×	×
$ambush_7$	×	×	0	-1	×	×	-1	2	×	×	3	3	×	×	1	0
$bamboo_1$	0	0	0	-1	3	3	3	-1	3	3	3	-1	-1	-1	-1	1
$bamboo_2$	-1	-1	-1	-1	3	-1	3	0	3	3	3	3	-1	1	1	1
$bandage_1$	-1	0	-1	-1	3	3	-1	2	3	3	3	3	-1	1	-1	1
$bandage_2$	0	0	0	-1	3	3	-1	-1	3	3	3	3	-1	1	-1	1
$cave_2$	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
$cave_4$	0	0	0	-1	-1	×	-1	-1	-1	3	0	-1	-1	0	-1	0
$market_2$	-1	-1	-1	-1	3	-1	3	0	-1	3	3	3	-1	1	-1	1
$market_5$	-1	-1	-1	×	×	×	×	×	0	0	0	0	×	0	0	×
$market_6$	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
$mountain_1$	-1	-1	-1	-1	2	3	-1	-1	2	-1	-1	-1	1	0	-1	0
$shaman_2$	0	0	0	-1	3	3	-1	-1	0	3	3	-1	-1	-1	3	0
$shaman_3$	0	-1	0	-1	-1	3	-1	-1	-1	3	0	0	-1	0	-1	0
$sleeping_1$	3	3	3	-1	3	3	3	-1	3	3	3	-1	3	3	3	3
$sleeping_2$	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
$temple_2$	-1	-1	-1	-1	-1	×	×	-1	-1	-1	-1	-1	-1	-1	-1	-1
$temple_3$	0	0	×	×	-1	-1	×	×	-1	0	×	×	×	-1	×	×

Table 3.2
PROFILE LABELS FOR EACH PAIR OF-CM.

ments, errors are arbitrarily large. In this case, none of the CMs is able to relate to the error. This is the case of the sequences which scored -1 for all CM and an OF methods, shown in figure 3.2. 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 3.16(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 3.16(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.

We can observe that different measures have different performances according to methods or sequences. For instance, 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. 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 performs worse for HS and NL, which are the OF methods that include more variability in sequences.

Thus, with the information given by the scores table, we consider as a proper OF-CM pairs to bound the OF error the following pairs: CLG-C_b, CLG-C_e, HS-C_b, HS-C_e, NL-C_e. The score table only provides a descriptive measure of the performance of a pair OF-CM. In order to choose a real OF-CM pair able to bound the OF error we require more sophisticated techniques such as ANOVA analysis, as developed in Chapter 4.

3.6 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 chapter, we have introduced a setting for categorizing confidence measures in terms of accuracy and capabilities for error bound prediction. We have introduced three graphical ways of exploring whether CM can provide a bound on OF error and a numeric categorization of each plot profile.

The presented tools have been validated on the Middlebury [48] and Sintel database [68], by four confidence measures with different grounds. The following interesting conclusions concerning quality plots as validation tools and capabilities of current CM for OF error bounding are derived from our experiment.

First of all, we have analyzed local capabilities of confidence measures for bounding the different types of OF error. We have also evaluated if current tools for confidence measure quality assessment agree with confidence measures bounding capabilities. Concerning the capabilities of existing CMs for OF error bounding, the following interesting points are derived from our analysis:

Energy (C_e) confidence measure detects when the functional is not properly minimized, and this usually happens at borders. In addition, this confidence measure only detects points where the computation does not correspond with the model assumptions, and thus it detects points that satisfy the model assumptions as reliable although they may not coincide with the ground truth.

Statistical (C_s) confidence measure detects if the computed flow is not coherent, and, thus, points having an OF either not regular or random. This implies that a constant OF would always be reliable, regardless of its agreement with ground truth.

Bootstrap (C_b) confidence measure detects if the model is unstable, that is, when small perturbations in the input data produce high variations in the output. Thus, it detects points that do not satisfy model assumptions like edges of object following different motions, and textureless regions.

Image Local Structure (C_k) confidence measure detects those pixels the image structure is not appropriate to solve optical flow, like textureless regions, and straight lines. However, textured regions may contain a lot of noise disturbing computations and, also, failing of OF assumptions is not considered.

We conclude that C_s is not the best suited for bounding OF error. Besides, C_k, C_b, C_e are able to bound a different kind of errors and, in fact, they provide complementary bounds.

Concerning existing methods for the evaluation of confidence measures, EPP better reflect non-decreasing profiles between CM and OF error and, thus, it is better suited for detecting CM unable to bound errors for a significant amount of cases. RAUC curves provide an improvement in the sense that the shown curve can assess the risk for each curve considered, however the space partition to obtain the plot is not invariant to transformations. Finally the SDP plot, has a good partition of the space that allows a good transformation scatter plot - descriptive plot. In addition, the SDP curve shows the risk at each CM percentile, which transforms the SDP curve into a useful tool able to properly assess a threshold on CM to assess error bound for OF. Thus, in the following chapters we will use only the SDP curve.

This chapter constitutes a first step in the use of statistical and probabilistic tools for the evaluation of the capabilities of CM for predicting OF error in decision support systems. In the following chapter (4) we present a solid methodology based on advanced statistical techniques to assess the capabilities of the confidence measures for error bounding.

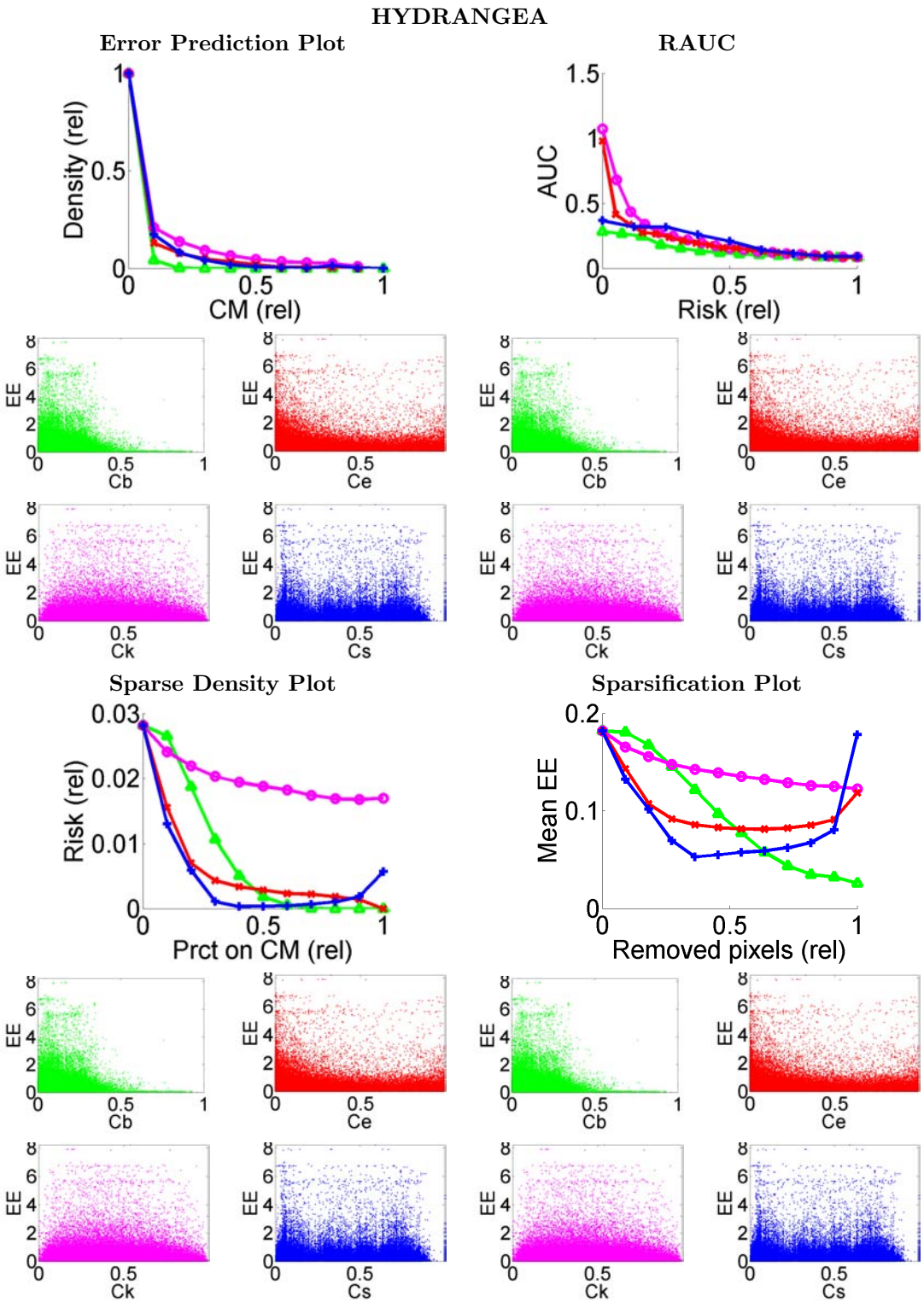


Figure 3.12: Descriptive plots for Hydrangea sequence.

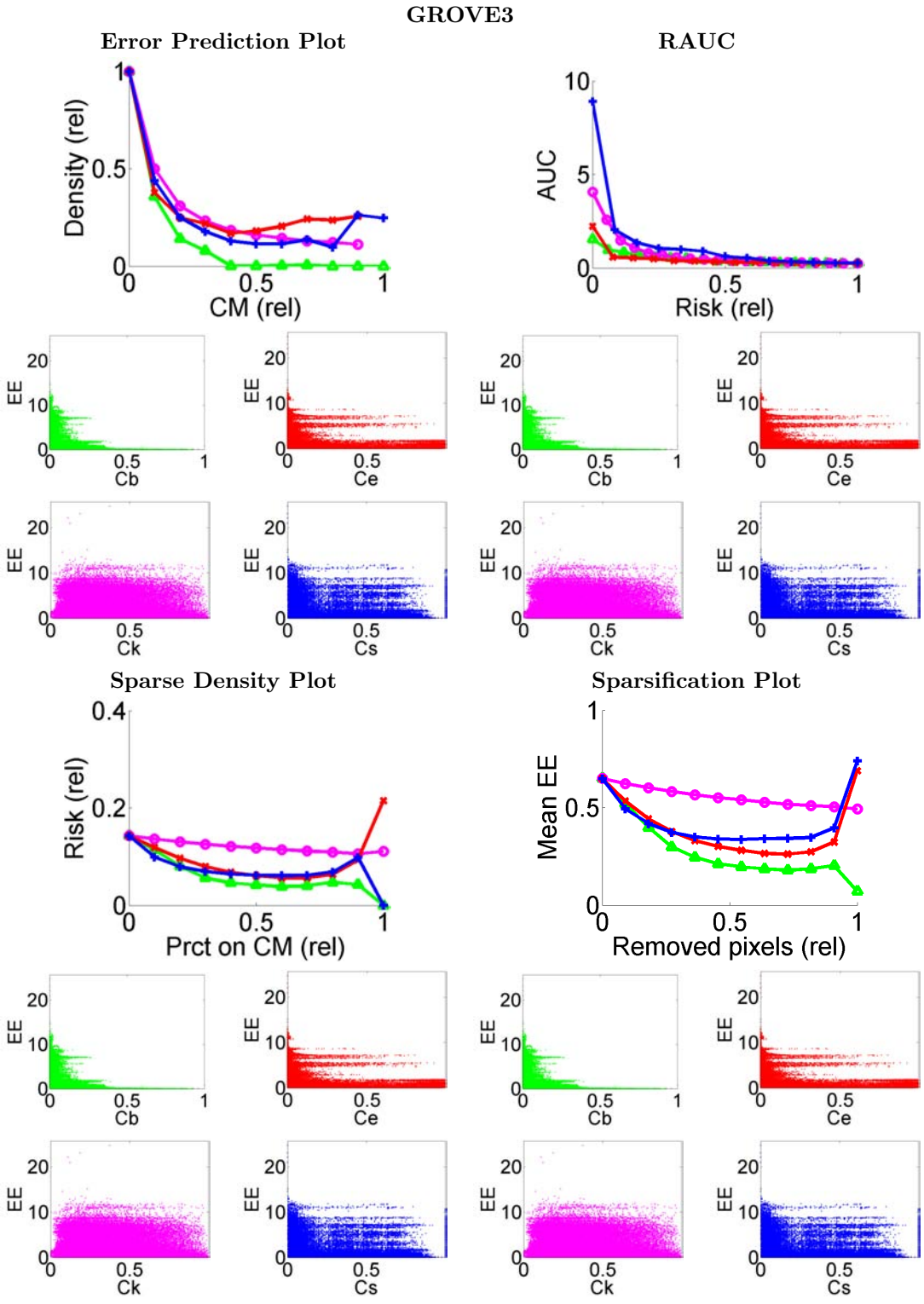


Figure 3.13: Descriptive plots for Grove3 sequence.

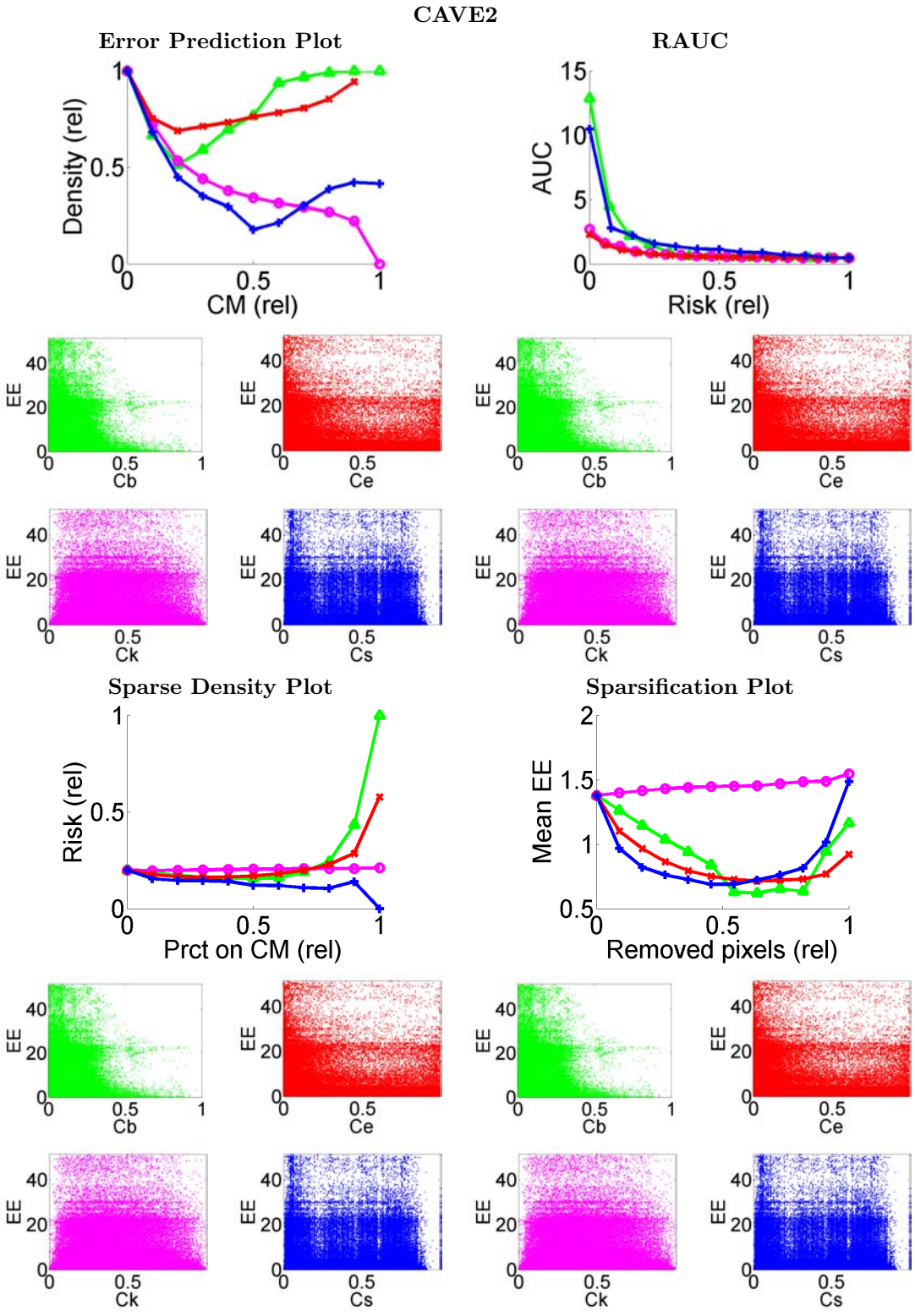


Figure 3.14: Descriptive plots for Cave2 sequence.

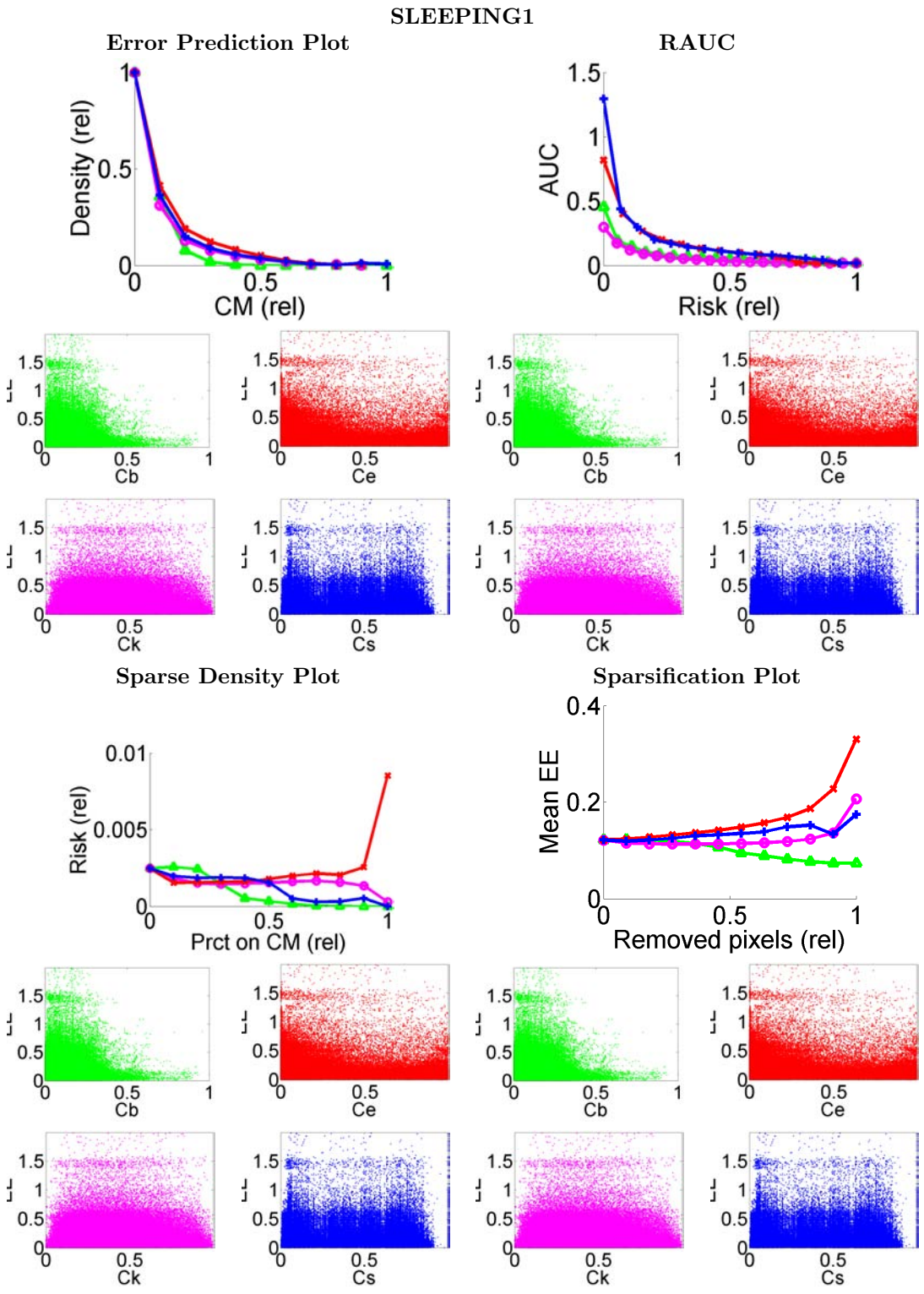
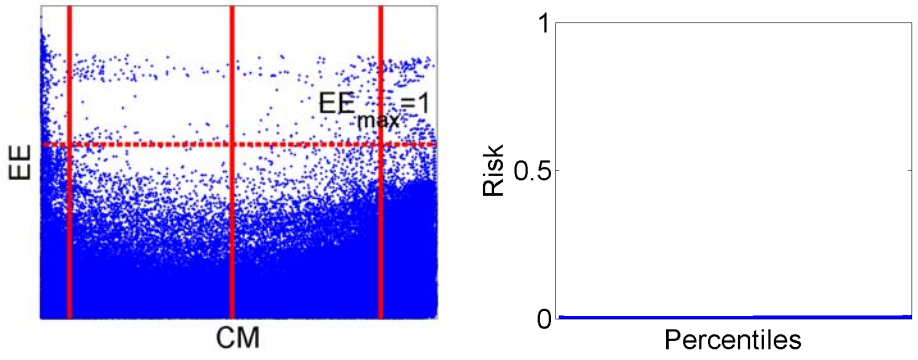
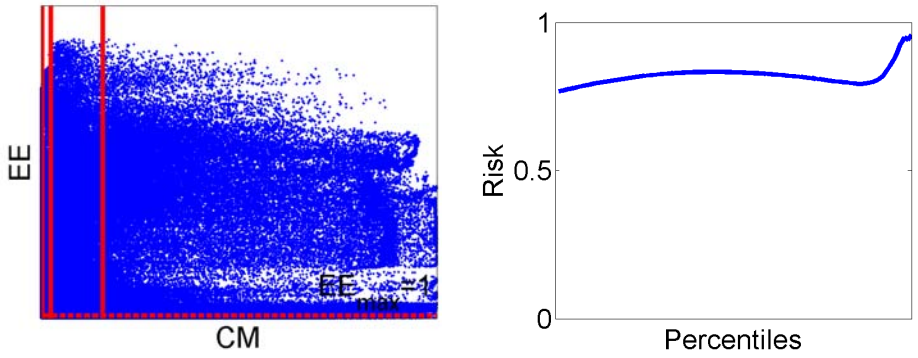


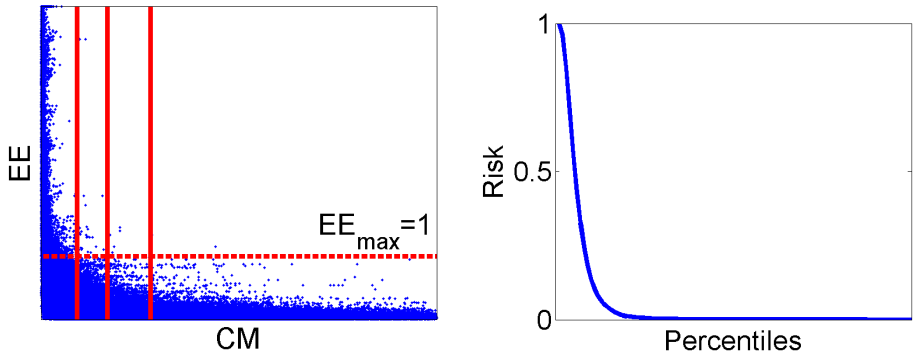
Figure 3.15: Descriptive plots for Sleeping1 sequence.



(a) Example of too good sequence.



(b) Example of too bad sequence.



(c) Example of predictable sequence.

Figure 3.16: 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.

Chapter 4

Statistical Framework for Comparison of Confidence Measure Predictive Value

In order to analyze the performance of different pairs OF-CM over a sequence, for every two consecutive frames, we compute the SDP curve. Thus, we have a sample of SDP curves and we can average them to obtain a curve, that we denote as $\mathcal{SDP}$. This curve can provide information about the general performance of the OF-CM pair. However, this curve is not enough to decide which pair OF-CM performs better, and this is because we do not have information about:

- Performance variability: we want to know how much do the results vary depending on the query, and also how significant is the difference between the performances of the different pairs OF-CM.
- Effect of factors of interest: we wonder which is the impact of a certain factor on the performance (sequences, optical flow methods, confidence measures).
- Generalization of the results: we need to know if a pair OF-CM can perform similarly in different conditions, in this case with sequences with different features.

In order to analyze deeply the performance of the different pairs OF-CM we use the statistical tools presented in chapter 2: confidence intervals, hypothesis tests, student's t-test and ANOVA. In our particular case, the factor is the combination of optical flow methods and the confidence measures, thus, the groups are the OF-CM pairs, and the quantitative response is a numerical score that evaluates if the confidence measure is capable of bounding the optical flow error. The objective of this test is to find significant differences among the pairs OF-CM and, in affirmative case to select the best parameter settings, that is, the best pair OF-CM that is able to bound the optical flow error by means of a multicomparison test. In the case of two-way ANOVA test, we consider two factors, which are sequences (F_S) and OF-CM pairs (F_P). The response variable is the same as for the one-way ANOVA test, a

numerical score that evaluates if the confidence measure is capable of bounding the optical flow error.

4.1 Statistical Framework Definition

In order to use a given CM in a decision support system, one should be able to provide a threshold on CM values that discards pixels prone to have high EE with a risk below the maximum risk allowed by the system. Such thresholds on CM values can be computed by means of a training set of frames representative enough of the decision support system sequences. We define a curve ($\mathcal{SDP}$) that provides a descriptive profile of SDP curves sampled across frames presenting similar appearance, dynamic conditions and resolution. For a given CM percentile $prct_{CM}$, let us consider its risk for a frame sampling of size N_{Fr} taken across a sequence type:

$$\rho_i = (\rho_i)_{i=1}^{N_{Fr}} := (\rho_i(prct_{CM}))_{i=1}^{N_{Fr}}$$

For each such percentile, let us consider the one-sided confidence interval [83] for the sample mean of the risk across the different frames ρ_i of a sequence. Provided that N_{Fr} is large enough, this interval can be computed at confidence level $1-\alpha$ from ρ_i sample mean, $\mu_{prct_{CM}}$, and variance, $\sigma_{prct_{CM}}$, as:

$$[0, \mu_{prct_{CM}} + t_{1-\alpha}^{N_{Fr}-1}\sigma_{prct_{CM}}] = [0, \Upsilon_{prct_{CM}}] \tag{4.1}$$

for $t_{1-\alpha}^{N_{Fr}-1}$ the value of a T-Student distribution with $N_{Fr}-1$ degrees of freedom having a cumulative probability equal to $1-\alpha$ [83].

The above punctual upper bound $\Upsilon_{prct_{CM}}$ taken across a fixed number of CM percentiles, provides a confidence curve that statistically describes SDP profiles across application sequence frames:

$$\mathcal{SDP} := \mathcal{SDP}(prct_{CM}) = (prct_{CM}, \Upsilon_{prct_{CM}})$$

The curve $\mathcal{SDP}$ provides an upper bound (see Section 4.2) for the error risk of new incoming sequences with conditions similar to the ones of the sample frames used to compute (4.1), provided that SDP variability across such a frame sample is as low as possible [84]. In this context, a most relevant quality feature of confidence measures would be a stable behavior of SDP across different sequences.

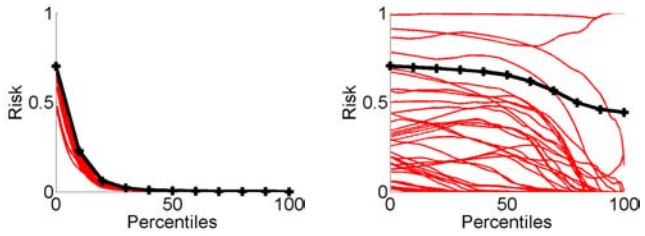
Therefore, the capability of a pair OF-CM for risk bounding should follow a two-stage cascade process. First $\mathcal{SDP}$ predictive value should be assessed and, then, for those OF-CM pairs with the highest predictive value, the quality of the bound provided by $\mathcal{SDP}$ should be determined.

In order to detect significant differences across several OF-CM pairs we use Analysis of Variance (ANOVA) [85] since it is a powerful statistical tool for detecting differences in performance across methodologies, as well as the impact of different factors or assumptions. We can apply ANOVA in case our data consists of one or several categorical explanatory variables (called factors) and a quantitative response of the variable. The variability analysis is defined as soon as the ANOVA quantitative

score and the different factors and methods are determined. Training data (individuals) are grouped according to such factors and differences among quantitative response group mean are computed. ANOVA provides a statistical way to decide if such differences are significant enough for a given confidence level α. In case of having more than one factor, ANOVA also detects any interaction across the different factors that might distort the analysis of results separately for each factor. If interaction across factors is significant, then the multiple ANOVA has to be re-designed as one factor ANOVA combining all factor groups into a single one. The ANOVA design (variable, individuals and factors) for each quality stage is defined as follows:

4.1.1 $\mathcal{SDP}$ Predictive Value

The curve $\mathcal{SDP}$ provides an effective bound for new cases if SDP variability across sequences is low. Figure 4.1 shows an illustrative example of the curve $\mathcal{SDP}$ computed over different samples of SDP curves. The SDP curves are depicted in red and the $\mathcal{SDP}$ one in black and crosses. Small variability indicates stable behavior of a pair OF-CM across different sequences or frames. In this case, the statistical rule (4.1) used to compute $\mathcal{SDP}$ guarantees a reliable upper bound of the error in the absence of ground truth for similar sequences with a confidence $1-\alpha$ (as illustrated in figure 4.1(a)). Otherwise, the model is not suitable for predicting sequences and frames not belonging to the ones used to compute $\mathcal{SDP}$ (as illustrated in figure 4.1(b)). Therefore, selecting those $OF-CM$ pairs that have $\mathcal{SDP}$ not presenting a significant variability is a first mandatory check for a further confident use. Significance in $\mathcal{SDP}$



(a) (b)

Figure 4.1: $\mathcal{SDP}$ (black line) predictive value for a sample of SDP curves (red) with small variability, (a), and with large and random variability, (b).

variability is checked using ANOVA as follows. Given a sampling of CM percentiles $prct^j_{CM} = \frac{j \cdot h}{N_{prct}}$, being h the sampling step and N_{prct} the number of percentiles, the variability of its $\mathcal{SDP}$ is approximated by the unbiased sample estimator:

$$\sigma^i_{\mathcal{SDP}} = \frac{1}{N_{prct}-1} \sum_{j=1}^{N_{prct}} \left(\rho_i(prct^j_{CM}) - \Upsilon^j_{prct_{CM}} \right)^2 \tag{4.2}$$

where i and j correspond to the frame and the percentile, respectively. We denote as $\sigma_{\mathcal{SDP}}$ the global variable depending on $\mathcal{SDP}$ that has the values of $\sigma^i_{\mathcal{SDP}}$ for all frames of a given sequence. In this context, the pair OF-CM best suited for a given application should be the one with the lowest $\sigma_{\mathcal{SDP}}$ for all application sequences.

Under the above considerations, ANOVA should explore the impact, if any, of sequence appearance and motion characteristics. Therefore, a 2-way ANOVA with factors defined by OF-CM pairs and sequences of the database are used to detect any interactions. This is a mandatory step if sequences are presumed to be heterogeneous (like Middlebury [48] or Sintel [68] benchmark databases). For each $\mathcal{SDP}$, a sampling of $\sigma_{\mathcal{SDP}}$ given by the N_{Fr} frames used to compute eq.(4.1) defines the individuals for each ANOVA factor group. If there is no significant interaction between OF-CM and sequence type, a 1-way ANOVA with OF-CM as factor and taking $\sigma_{\mathcal{SDP}}$ percentile sampling for all sequences as individuals serves to select those OF-CM pairs best suited for the application. In case of interaction, a 1-way ANOVA with the combined factor (OF-CM)-Seq taken over $\sigma_{\mathcal{SDP}}$ percentile sampling should be applied to determine which sequences have homogeneous features for OF-CM performance.

This ANOVA selects those OF-CM pairs that have a significantly large variability across SDP profiles and, thus, such pairs should be dropped from the further analysis.

4.1.2 $\mathcal{SDP}$ Bound Quality

For those pairs OF-CM and sequences such that $\mathcal{SDP}$ variability is significantly low, $\mathcal{SDP}$ is a good descriptor of the relation between OF error and CM. Two aspects determine the quality of $\mathcal{SDP}$ bound: a good correlation with OF error and a minimum risk for the bounded pixels.

The correlation to OF error is directly assessed by the $\mathcal{SDP}$ label assigned using the conditions given in Section 4.1.1. Given that categorization of SDP curve profiles with labels $\{2, 1\}$ strongly depends on each application's specific requirements, all $\mathcal{SDP}$ having a label greater than 0 will be considered to have a suitable decreasing profile.

Like in subsection 4.1.1, the amount of risk is assessed by defining a variable depending on $\mathcal{SDP}$ that can be compared across sequences and OF-CM pairs using ANOVA. In this case, the ANOVA variable, noted by $\Upsilon_{\mathcal{SDP}}$, is the average of SDP values computed across a sampling of $prct_{CM}$:

$$\Upsilon^i_{\mathcal{SDP}} := \frac{1}{N_{prct}} \sum_j \rho^i(prct^j_{CM}) \tag{4.3}$$

for $prct^j_{CM} = \frac{j \cdot h}{N_{prct}}$ and the subscript i denoting each individual frame of a given sequence. In this case, a 1-way ANOVA is applied having as factors either the selected OF-CM pairs or the combined factor (OF-CM)-Seq. The choice will depend on whether the 2-way ANOVA carried out in the first quality stage detects interactions or not.

The OF-CM best suited for the application will be the pair having the best predictive power, an $\mathcal{SDP}$ label above 0 and the least significant risk detected in this second stage ANOVA.

4.2 $\mathcal{SDP}$ Applicability

The $\mathcal{SDP}$ curve learned at confidence $1-\alpha$ can be used to assess OF for similar sequences without ground truth in two aspects:

1. **Confident Pixel Discarding.** The inverse of the curve $\mathcal{SDP}$ allows selecting for each new incoming frame the set of pixels such that their risk is under a given maximum risk allowed by the application with confidence $1-\alpha$.

2. **Confident Risk Assessment.** The curve $\mathcal{SDP}$ itself assesses the expected risk for a given percentage of image pixels with confidence $1-\alpha$. This allows the computation of outliers over the minimum number of pixels required for further computations.

4.2.1 Confident Pixel Discarding

Let ρ_{Mx} be a maximum risk required in our decision support system and consider the intersection of the horizontal line $\rho \equiv \rho_{Mx}$ with the curve $\mathcal{SDP}$.Such intersection point is given by a CM percentile and corresponds to the inverse of the function $\Upsilon_{prct_{CM}}$ evaluated at ρ_{Mx}, $\Upsilon^{-1}_{prct_{CM}}(\rho_{Mx})$. We would like to note that the percentile $\Upsilon^{-1}_{prct_{CM}}(\rho_{Mx})$ should be large enough to allow further computations in the decision support system.

The percentile $\Upsilon^{-1}_{prct_{CM}}(\rho_{Mx})$ provides the actual threshold over CM values by means of CM distribution function, $F_{CM}(cm) := P(CM \geq cm)$. To be precise, the inverse:

$$F^{-1}(\Upsilon^{-1}_{prct_{CM}}(\rho_{Mx})) \tag{4.4}$$

is the threshold on CM values such that $100(1-\rho_{Mx})\%$ of the image pixels with CM over (4.4) have an error under EE_{max} in $100(1-\alpha)\%$ of the frames. In case CM had a different scale for each frame, the threshold (4.4) would be computed for each new frame using the inverse of the empirical distribution computed over its pixels. Otherwise, (4.4) would be common to any frame and could be directly computed from the distribution function of the training set.

At this point, it is worth explaining the meaning of the confidence α used to compute $\mathcal{SDP}$ from an application point of view. By the properties of confidence intervals, we have that the probability that a frame has a risk above ρ_{Mx} is approximately the confidence α used to compute $\mathcal{SDP}$. This implies that for a new sequence the proportion of failing frames is approximately α. Therefore, in practice, the number of failing frames depends on the number of sequence frames as well as on the confidence α and can be computed using confidence intervals. We will call such expected number of failing frames Empirical Confidence (EC).

By definition, the failing cases follow a binomial distribution of probability $p = \alpha$, $B(n,p) = B(n_{Fr}, \alpha)$, for n_{Fr} the number of sequence frames:

$$P(B(n_{Fr},\alpha) = k) = \binom{n_{Fr}}{k} \alpha^k (1-\alpha)^{n_{Fr}-k} \tag{4.5}$$

Therefore the number of expected failures [86] can be estimated with confidence $1-\alpha$ as the smallest value such that a binomial $B(n_{Fr},\alpha)$ distribution function equals

$1-\alpha$:

$$P(B(n_{Fr},\alpha) \leq EC) = \sum_{k=1}^{EC} \binom{n_{Fr}}{k} \alpha^k (1-\alpha)^{n_{Fr}-k} = 1-\alpha \tag{4.6}$$

Such a value is easily computed as the inverse of the binomial distribution function.

4.2.2 Confident Risk Assessment.

The values of the curve $\mathcal{SDP}$ return the expected percentage of outliers without bounded error for each percentage of image pixels selected according to CM percentiles. This is useful in case our application requires a minimum amount of pixels $prct_{Mm}$ for a reliable performance of the decision support system. In such case, $\mathcal{SDP}$ value at $prct_{Mm}$:

$$\Upsilon(prct_{Mm}) \tag{4.7}$$

directly assesses the expected risk with confidence $1-\alpha$.

As before, confidence should be interpreted in terms of EC. That is, the number of frames such that the outliers with an error over EE_{max} is less than (4.7) is approximately $100(1-\alpha)\%$ of the total number of frames in the sequence. This number can be actually computed using the inverse of the distribution function of a binomial $B(n_{Fr}, 1-\alpha)$ as the minimum number such that:

$$P(B(n_{Fr},1-\alpha) \leq EC) = \sum_{k=1}^{EC} \binom{n_{Fr}}{k} \alpha^{n_{Fr}-k} k (1-\alpha)^k = 1-\alpha \tag{4.8}$$

4.3 Experimental Setup

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 cover as much methods and sequence features as possible, we have chosen the Sintel database [68], four representative OF methods, and four confidence measures. These settings give $4 \times 4 = 16$ possible OF-CM pairs.

Database. In order to get consistent statistics, sequences should have ground truth and more than 30 frames (30 frames for training and the rest for testing). The Sintel database [68] fulfils this condition so we have selected 17 sequences with ground truth and at least 40 frames. Besides, it contains sequences with large motion, specular reflection, motion blur, defocus blur and atmospheric effects, so that it covers a complete battery of sequence features.

Optical flow algorithms. Our framework has been applied to classic and state of the art representative OF methods [1]. The classic formulations are:

[1] Using the free source code from [82], [40] and [67].

	Ck				Cb				Ce				Cs			
	CLG	HS	NL	Corr	CLG	HS	NL	Corr	CLG	HS	NL	Corr	CLG	HS	NL	Corr
$alley_1$	0.03	0.05	0.03	0.02	0.00	0.01	0.01	0.02	0.00	0.01	0.00	0.01	0.01	0.02	0.01	0.01
$alley_2$	0.02	0.05	0.01	0.02	0.01	0.01	0.01	0.02	0.01	0.01	0.00	0.03	0.01	0.01	0.01	0.00
$ambush_2$	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
$ambush_4$	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
$ambush_5$	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
$ambush_6$	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
$ambush_7$	×	×	0.00	×	×	×	×	0.01	×	×	0.00	0.01	×	×	0.00	0.02
$bamboo_1$	0.05	0.07	0.05	0.05	0.01	0.01	0.01	0.06	0.01	0.01	0.01	0.04	0.02	0.02	0.02	0.02
$bamboo_2$	0.06	0.08	0.05	0.05	0.00	0.01	0.01	0.03	0.00	0.01	0.00	0.01	0.01	0.01	0.01	0.01
$bandage_1$	0.11	0.16	0.07	0.06	0.02	0.04	0.06	0.05	0.04	0.04	0.01	0.03	0.04	0.05	0.02	0.03
$bandage_2$	0.07	0.11	0.07	0.05	0.02	0.03	0.04	0.04	0.02	0.02	0.02	0.03	0.03	0.05	0.03	0.02
$cave_2$	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
$cave_4$	×	×	0.17	×	×	×	×	×	0.22	0.07	0.07	0.19	0.20	0.11	0.08	0.08
$market_2$	0.09	0.11	0.07	0.09	0.01	0.03	0.01	0.03	0.02	0.01	0.01	0.02	0.02	0.03	0.01	0.02
$market_5$	×	×	×	×	×	×	×	×	×	×	0.34	×	×	×	0.33	×
$market_6$	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
$mountain_1$	0.11	0.43	0.11	0.20	0.06	0.19	0.13	0.22	0.06	0.22	0.09	0.15	0.05	0.35	0.09	0.15
$shaman_2$	0.03	0.06	0.04	0.06	0.01	0.02	0.03	0.06	0.02	0.02	0.02	0.05	0.01	0.05	0.02	0.03
$shaman_3$	0.11	0.13	0.05	0.04	×	0.04	0.06	0.05	×	0.03	0.02	0.04	0.08	0.08	0.03	0.03
$sleeping_1$	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00
$sleeping_2$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$temple_2$	0.17	0.23	0.16	0.27	0.11	×	×	×	0.14	0.07	0.09	0.25	0.13	0.18	0.10	0.23
$temple_3$	0.86	0.88	×	×	0.96	0.86	×	×	×	×	×	×	×	0.88	×	×

Table 4.1
$\mathcal{SDP}$ BOUND QUALITY FOR THE SINTEL DATABASE.

- Combined Local-Global (CLG) [38], which uses a Lucas-Kanade data term [33] with an L^2 norm smoothness term.
- Horn-Schunck (HS) [35], which uses OF brightness constancy equation with an L^2 norm smoothness term.

while the state of the art are:

- Classic-NL (NL) [40], which is a total variation method that uses the L^1 norm to combine OF brightness constancy assumption with the smoothness term.
- Correlation method ($Corr$) [67], which uses an L^2 data term based on the correlation transform of the images with an L^1 regularity term based on bilateral filtering.

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 [31], we have selected the condition number of the data-term system defined in [79].
- Energy (C_e): The confidence measure is computed by evaluating the flow field over the functional, as described in [49].
- 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 [87].
- Bootstrap (C_b): It measures OF variability with respect to perturbations in the model [53].

For the experiments, all CM are assumed to take values in the interval $[0, 1]$, 0 meaning low confidence and 1 high confidence. On the one hand, C_k and C_s are already in the range due to their definition. On the other hand, C_e and C_b need to be normalized. This normalization can be a global one given by CM definition (such as C_e) or an empirical one computed over a sample (such as C_b).

Experiments. Three experiments are presented. First, the whole framework is applied to a training set of sequences to select the best suited OF-CM pairs. Second, the $\mathcal{SDP}$ bounds for the selected pairs are applied to a test set of sequences to show the actual predictive power of the computed bounds. The significance of all tests and statistics is $\alpha = 0.05$. Finally, we show an example of the applicability of the presented framework.

For the first experiment, the ANOVA variables $\sigma_{\mathcal{SDP}}$, $\Upsilon_{\mathcal{SDP}}$ have been computed over the uniform sampling of CM percentile $prct^j_{CM} = \frac{j0.1}{N_{prct}} = \{0, 0.1, \ldots, 1\}$ and using 30 random frames for each sequence. The maximum allowed error for computation of SDP has been set to $EE_{max} = 1$.

For the second experiment, we have assessed the actual predictive capabilities of each OF-CM pair in 2 aspects: prediction power and percentage of discarded pixels. The ideal situation would be a high prediction power with the lowest percentage of discarded pixels.

For each OF-CM, we have considered the triplets OF-CM-Seq that according to the first experiment can be used to bound the error. For each such triplet, we have used the training curve $\mathcal{SDP}$ to compute the threshold on CM given by (4.4) for a risk $\rho_{Mx} = 0.05$. The confidence interval for the percentile $\Upsilon^{-1}_{prct_{CM}}(\rho_{Mx})$ used to compute (4.4) provides the average percentage of discarded pixels. To compute prediction power, pixels on the test frames have been discarded according to (4.4) and the number of test frames with a risk over $\rho_{Mx} = 0.05$, which we call Real Empirical Confidence (REC) has been computed. Under the hypothesis that CM can bound OF error, the expected REC should be less than EC given by (4.6). Since the number of test frames is $n_{Fr} = 17$ and the confidence is $\alpha = 0.05$, the number of failing frames given by (4.6) should be less or equal than 3. A left tailed t-test for an average REC under 3 serves to check the prediction power of each OF-CM pair. A $p - val$ under $\alpha = 0.05$ rejects the null hypothesis and ensures with $1 - \alpha = 0.95$ of confidence that the OF-CM pair has an average risk under $\rho_{Mx} = 0.05$.

The last experiment shows an example of the points that we discard after applying the presented framework.

4.4 Results

4.4.1 Exp1: Framework Training.

The 2-way ANOVA over OF-CM and sequence factors gives significance for each factor (with $p-val < 10^{-16}$) as well as interaction ($p-val < 10^{-16}$). Interaction is mainly attributed to the heterogeneity of the Sintel database, which contains sequences with several artifacts (blurring, illumination changes, large displacements, etc.) that influences the error bounding properties of the considered OF-CM pairs. Therefore, the remaining 1-way ANOVA is applied to the combined factor OF-CM-Seq, which results in a $4 \times 4 \times 17 = 272$ groups to be analyzed.

Table 4.1 shows the bound quality assessment for those OF-CM-Seq triplets that had the least significant variability across sequence frames. We show the average risk for each triplet with those ones having the least significant risk according to ANOVA highlighted in blue. Crosses indicate that the triplet had a significant large variability and, thus, OF errors can not be predicted.

There are four sequences ($cave_2$, $ambush_4$, $ambush_5$ and $market_6$) that could not be predicted by any pair due to a significantly large variability across SDP curves. Such variability mainly arises in case frames have highly heterogeneous motion and intensity patterns producing good and bad SDP profiles. For instance, sequence $ambush_4$ starts with small displacements, and suddenly from frame 3 to 4 a huge object appears on the image and the sequence has large displacements and blurring from there until frame 13 approximately, where sequence again has sharpen objects and smaller displacements. Such abrupt changes across the sequence split SDP profiles into two main groups, which introduce a large variability.

4.4.2 Exp2: Framework Testing.

For those triplets that have been selected as good candidates for error bounding (highlighted in blue in table 4.1), we have checked their actual predictive power on the test frames. Results for each $OF-CM$ pair are reported in Table 4.2. We report the CI interval for REC (Real Empirical Confidence) taken over all candidates, labelled CI_{REC}, the p-value, labelled p_{REC}, of a left-tailed t-test with null hypothesis checking wether the average REC is above 3 and the CI for the average percentage of discarded pixels, labelled $CI_{\Upsilon^{-1}_{prct_{CM}}}$.

The measure C_s is, by far, the worst suited for OF bounding as, with $p_{REC} > 0.7$, it has a risk above the expected 3 for all OF methods. In fact, the CI for the average REC predicts an increase in risk up to 8 frames at least, which represents almost half (47%) of the test frames. The measure C_k has the highest prediction power, with 3 OF methods (HS, NL and Corr) achieving the expected REC in average. However this is at the cost of the highest pixel discarding rate, with intervals $CI_{\Upsilon^{-1}_{prct_{CM}}}$ indicating that all image pixels could be removed. This fact, invalidates C_k for most decision support system, since usually a further use of OF within the application computations is required. The measure C_b also achieves an average REC under 3 for 3 OF methods (CLG,HS and Corr). Although, it has a percentage of discarded pixels acceptable only for the classic CLG and HS. Finally, the measure C_e presents a good pixel

	CI_{REC}	p_{REC}	$CI_{\Upsilon^{-1}_{prct_{CM}}}$
CLG-C_k	(0, 4.0)	0.17	(50 %, 103 %)
HS-C_k	(0, 2.8)	0.02	(55 %, 109 %)
NL-C_k	(.3, 2.0)	7e-6	(57 %, 103 %)
Corr-C_k	(.2, 2.0)	3e-5	(41 %, 104 %)
CLG-C_b	(.5, 2.9)	0.02	(12 %, 56 %)
HS-C_b	(.1, 1.6)	1e-4	(20 %, 67 %)
NL-C_b	(.2, 4.6)	0.3	(18 %, 75 %)
Corr-C_b	(1, 2.9)	0.02	(45 %, 102 %)
CLG-C_e	(.6,5.5)	0.5	(13 %,60 %)
HS-C_e	(.7, 4.0)	0.2	(17 %, 53 %)
NL-C_e	(1, 4.8)	0.6	(17 %, 58 %)
Corr-C_e	(0,1.3)	3e-6	(26 %, 72 %)
CLG-C_s	(.6, 7.6)	0.7	(18 %, 61 %)
HS-C_s	(3, 12)	1.0	(35 %, 86 %)
NL-C_s	(2, 8.9)	0.9	(22 %, 64 %)
Corr-C_s	(2 ,8.8)	0.9	(25 %, 71 %)

Table 4.2
$\mathcal{SDP}$ BOUND TESTING. STATISTICAL SUMMARY

discarding rate for all OF methods, but only achieves the expected REC for Corr. The capabilities of CM for OF error bounding are summarized in Table4.3.

The results shown in table 4.3, indicate that measures based on either local image structure [31] or local motion regularity [87] are not the best suited for predicting OF error risk, at least for the considered OF methods. Energy-based [49] and bootstrap [53] 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. In this context, the best candidates to predict error risk are CLG-C_b, HS-C_e and NL-C_e.

It is worth noticing that none of the measures was well posed for bounding $Corr$ error. This might be attributed to a high specialized formulation. Our framework assesses wether a generalist algorithm makes significant mistakes for a given application. The behavior of generalist algorithms, such as classic ones CLG, HS, can be easily predicted using a single measure but are prone to give less accurate results (as algorithm rankings like the ones found in http://sintel.is.tue.mpg.de/ indicate). In case no general algorithm could be selected, the approach proposed in [54] could be used to select, for each pixel, the specialist algorithm best suited for the pixel particular appearance and temporal features.

This selects the pair Corr-C_e as the best suited with an average proportion of frames with unbounded error within $CI_{REC} = (-0.5, 1.3)$, 90% of sequences achieving the expected risk and at most 72% of discarded pixels.

	Prediction	Discarding
CLG-C_k	×	×
HS-C_k	✓	×
NL-C_k	✓	×
Corr-C_k	✓	×
CLG-C_b	✓	✓
HS-C_b	✓	✓
NL-C_b	×	✓
Corr-C_b	✓	×
CLG-C_e	×	✓
HS-C_e	×	✓
NL-C_e	×	✓
Corr-C_e	✓	✓
CLG-C_s	×	✓
HS-C_s	×	×
NL-C_s	×	✓
Corr-C_s	×	✓

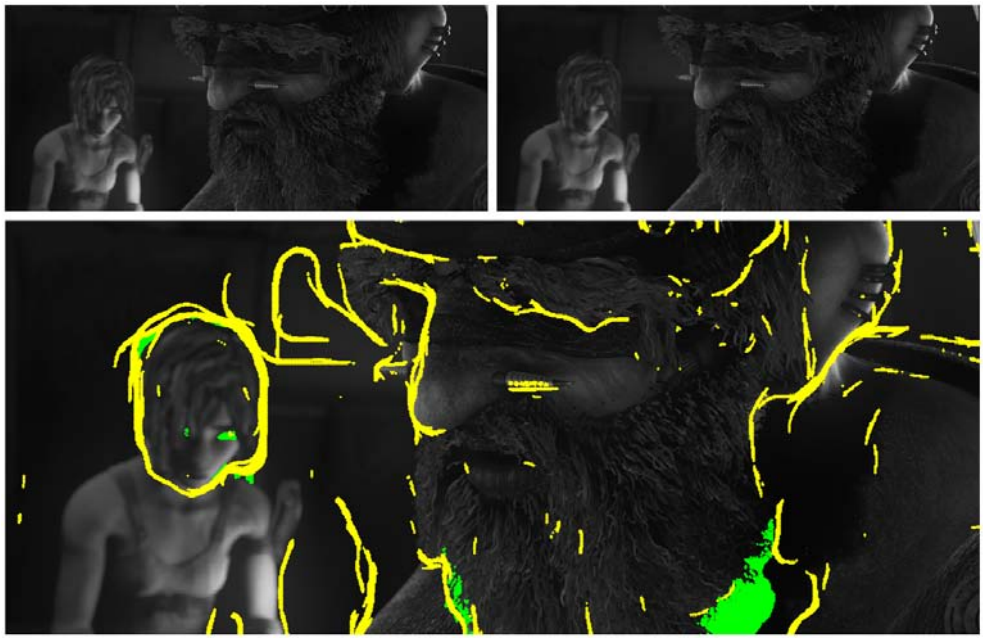
Table 4.3
Capabilities of current CMs for OF error bounding

4.4.3 Exp 3: Applicability of the presented framework

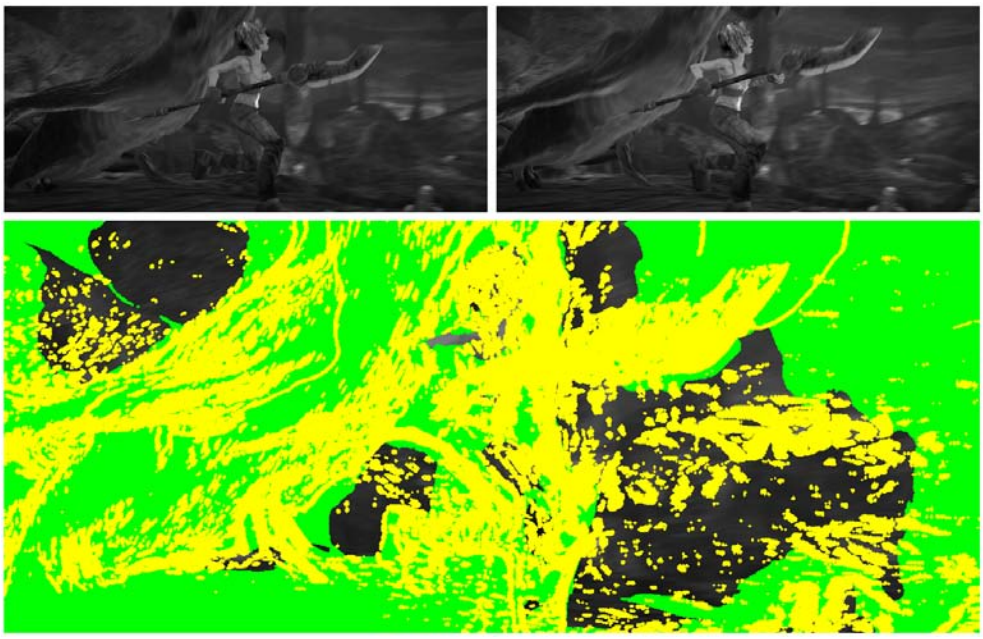
Figure 4.2 shows an example of the points that we discard after applying the presented framework. On the first row, two consecutive frames of the sequence. On the second row, depicted in yellow, the points discarded by the application using a risk of 0.05 and, in green, remaining points with error larger than EE_{max}. Figure 4.2(a) shows a screen shot of sequence predictable using the pair NL-Ce. Figure 4.2(b) shows a screen shot of sequence ($cave_2$) which could not be predicted by any pair. For each case, we also report the percentage of discarded pixels (PR) and the percentage of pixels that could not be bounded (PU). For the predictable case, PU is under the 5% of allowed risk, while PU significantly increases to almost 50% for the unpredictable one.

4.5 Conclusions

This chapter provides the definition of a novel framework based in the use of statistical and probabilistic tools for the evaluation of the capabilities of CM for predicting OF error in decision support systems. As any other statistical approach, our methodology can be applied as far as the training sequences are representative enough of the the dynamical appearance of application frames. This might require either a large data set capturing the variability of frames appearance or controlled acquisition conditions. The latter are frequently achieved in protocols used in medical imaging or the vision systems of industrial applications.



(a) NL-Ce-$shaman_2$. Fr 42-43. PR=3.5% PU=0.9%



(b) CLG-Ce-$cave_2$. Fr 17-18. PR=31.0% PU=46.9%

Figure 4.2: Pixel discarding for a predictable (a) and non predictable (b) case. On top, the two consecutive frames used for the visualization are shown (Fr frames number). On bottom, the discarded pixels are depicted in yellow (PR percentage of discarded pixels). From remaining pixels, pixels with error larger than $EE_{max} = 1$ are depicted in green (PU percentage of unbounded pixels from the remaining ones).

Chapter 5

Application to medical imaging

In this chapter we apply the presented framework to improve clinical decision support systems. First of all, we use the presented framework in a manually annotated dataset of bronchoscopy guidance to improve the navigation inside the bronchial tree. Second, we analyze the impact of different image acquisition artifacts in cardiac diagnose systems, such as noise and decay. In order to do so we use a synthetic dataset reflecting those artifacts and ANOVA.

5.1 Confident Tracking of Anatomical Structures in Video-Bronchoscopy

In order to illustrate the applicability of the framework presented in Chapter 4, we use $\mathcal{SDP}$ to discard pixels prone to present large OF error in the context of tracking tracheal structures in videobronchoscopy explorations. Bronchoscopy is an endoscopy procedure used to visualize the inside of the airways for diagnostic and therapeutic purposes. The bronchoscope is inserted into the airways and allows the physician to examine the patient's airways for foreign bodies, bleeding, tumors or inflammation.

Bronchoscopists have to navigate through the bronchial tree until the lesion is reached. This task is not trivial due to the fact that lesions might be located at a very distant bronchial tree ramifications. Thus, physicians require a navigation system which help them to reach such lesions by providing 3D measures of the structures that appear in the scene. This navigation performed efficiently would provide a more efficient performance of the physician during the intervention, it would potentially help the doctor to locate the endoscope inside an specific organ and to monitor the patient's lesion. Nowadays this is not performed automatically, leading to an additional cost associated to the repetition of the interventions. Navigation systems based on computer vision techniques try to find corresponding points between two consecutive frames by using feature points or detecting anatomical structures among others. As an example, a navigation system tracks anatomical structures such as tracheal rings to find correspondences to estimate inter-frames movements. These correspondences can be found by computing optical flow techniques on tracheal rings in order

to detect their movement. Unfortunately, there are always errors in the optical flow estimation. Such errors can ruin the tracking of the rings and the movement estimation. In order to detect values of optical flow with high errors (outliers), confidence measures and statistical tools explained in Chapter 4. Points marked as outliers by our statistical tools are removed from the tracking process and then, the movement estimation improves drastically the results of the tracheal ring tracking. Even the error between iterations of a system is small, since the navigation problem is solved as a tracking problem, the error is accumulative and it becomes higher at each iteration. Hence the importance of removing correspondence with high errors is critical in any navigation system.

Nowadays, camera pose estimation is an active topic of research and it is useful for autonomous driving systems, object recognition systems or augmented reality applications. A common approach is to use key-points such as SIFT or SURF in order to estimate the inter-frame movement. Thus, SIFT descriptors are computed in both images and are used to find correspondences or matches by searching the most similar descriptors between the images. These matches can be already used to estimate the fundamental matrix (F) which is a matrix that allows to retrieve camera poses. The problem is that this estimation process is very sensitive to noise. For that reason, we need to remove outliers in order to increase the precision of the estimation. The most accepted method to remove outliers from a set of data is RANdom SAmple Consensus [88] (RANSAC). RANSAC can deal with a significant number of outliers present in the set of matches. The basic idea of the algorithm is to compute the fundamental matrix F using seven randomly selected matches (least number of points that the estimation method needs to solve the problem). The next step of RANSAC is to verify epipolar constraints with the estimated Fundamental matrix for all the matches. If the euclidean distance between points and their corresponding epipolar lines is smaller than a threshold the match is consider as an inlier. This process is repeated a sufficient number of iteration N. At the end, the fundamental matrix than produced more inliers is the fundamental returned by the method. The camera pose can be extracted from the result matrix of RANSAC.

The $\mathcal{SDP}$ curve has been learned for C_e and $Corr$ using 30 frames sampled from 5 representative sequences courtesy from Hospital de Bellvitge (for more information see Chapter 2). The maximum error for computing the sample SDP_i, $i = 1, \ldots, N_{Fr} = 30$, is set, as before, to $EE_{max} = 1$ and the confidence for $\mathcal{SDP}$ is also $\alpha_{\mathcal{SDP}} = 0.05$. In this case, errors are computed as the distance of the tracked structures to manually traced tracheal rings.

The learned curve $\mathcal{SDP}$ together with the training curves are shown in fig.5.1. The variability across frames is a bit higher than for synthetic sequences, but low enough with and average $\sigma^{SDP} \in []$. The increase in risk for high CM percentiles observed in two cases is due to isolated pixels and is quite often in real applications mainly due to a suboptimal definition of the manual ground truth. Still, the average AUC^{SDP} is within $[0.0696, 0.1457]$ with confidence $1 - \alpha = 0.95$ and compares to the best figures obtained for the Sintel predictable cases (see Table4.1).

For this application, $\mathcal{SDP}$ is used to evaluate if the expected risk after removing a maximum number of points is acceptable for camera pose estimation. To ensure that enough pixels are kept, we set $prct_{Mm} = 40\%$ to compute the expected risk

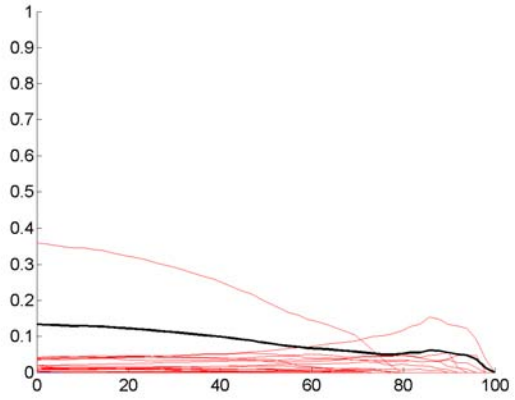


Figure 5.1: Confident Tracking in videobronchoscopy, $\mathcal{SDP}$ (in black) computed from 30 training SDP_i curves (in red).

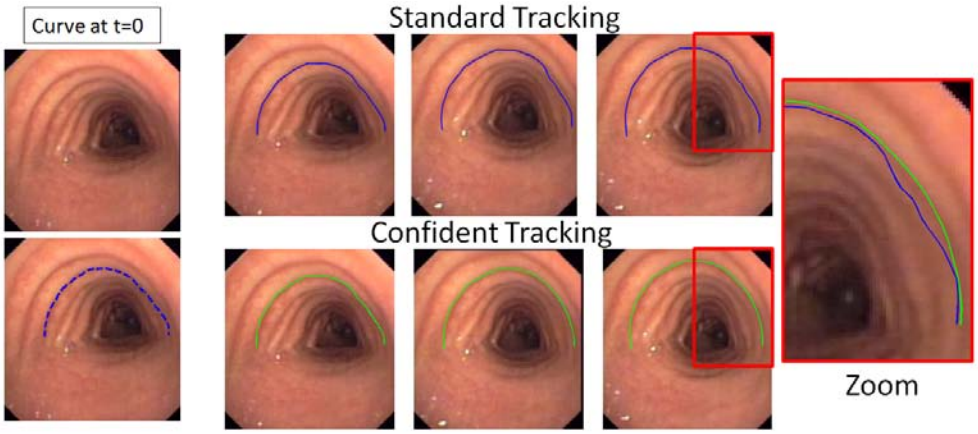


Figure 5.2: Confident Tracking in videobronchoscopy. Comparison between standard tracking and confident tracking discarding outliers.

$\Upsilon(prct_{Mm}) = \Upsilon(40) = 0.2778$. This implies that approximately at most 1/10 of the selected pixels might be outliers, which, according to the literature, is good enough for SLAM reconstruction and camera pose estimation.

To validate $\mathcal{SDP}$ risk bounds, we have conducted 2 experiments (summarized in Table 5.1) on 4 sequences different from the training ones and having between 17 and 40 frames each (n_{Fr} in Table 5.1).

On one hand, we have computed (like the experiment reported in Chapter 4) the actual REC to check that is under the expected EC given by (4.8). For each sequence, we have also computed the confidence interval for the average risk, CI_Υ , and checked it is under $\mu(\Upsilon) = 0.2778$ using a one tailed t-test with null hypothesis

	n_{Fr}	EC	REC	CI_{Υ}	p_{Υ}	$CI_{EE_T - EE_{CT}}$	$p_{EE_T - EE_{CT}}$
Seq_1	40	35	35	$(-\infty, 0.1896]$	4.0e-009	$[0.63, 1.44]$	1.3373e-008
Seq_2	30	26	28	$(-\infty, 0.1664]$	2.1e-011	$[0.35, 0.55]$	5.1135e-013
Seq_3	30	26	30	$(-\infty, 0.1452]$	2.9e-013	$[0.03, 0.07]$	2.8234e-007
Seq_4	17	14	16	$(-\infty, 0.0601]$	8.9e-014	$[0.00, 0.09]$	0.0283

Table 5.1
TRACKING OF ANATOMICAL STRUCTURES IN VIDEOBRONCHOSCOPY. STATISTICAL SUMMARY

$H_0 : \mu(\Upsilon) > 0.2778$ at $\alpha = 0.05$. We note that p-values, labelled p_{Υ} in Table 5.1, should be under $\alpha = 0.05$ to reject H_0 and ensure that $\mu(\Upsilon)$ is under the risk predicted by $\mathcal{SDP}$. The numbers reported in Table 5.1, show that the frame risk is under $\mathcal{SDP}$ prediction for all sequences ($p_{\Upsilon} \leq e-9$) and $REC \geq EC$. This validates our framework as predictive tool.

On the other hand, we have computed the tracking error across sequence frames obtained using all points and only the confident ones. Confident point tracking is computed by selecting, for each frame, points on the current tracked ring having CM above the percentile $prct_{Mm} = 40\%$. Such set is updated using the computed OF and their positions are used to interpolate the positions of the non-confident points to restore a whole ring. Tracking errors are given by distances to manually annotated rings and significant differences between errors for standard (labelled EE_T) and confident tracking (labelled EE_{CT}) are assessed using a one-tailed t-test for paired data at $\alpha = 0.05$. We set the null hypothesis to $H_0 : \mu(EE_T - EE_{CT}) < 0$, so that rejecting the test ($p_{EE_T - EE_{CT}} < 0.05$) shows that confident tracking has smaller average errors than the standard tracking. Again, for all sequences the null hypothesis is rejected and confident intervals for $\mu(EE_T - EE_{CT})$ show that differences can be large for some sequences (Seq_1).

Figure 5.2 shows an example of the benefits of using confident measures to discard potential outliers in tracheal ring tracking. The most-left image shows a ring manually traced (dashed blue line) on an initial sequence frame, top images its standard tracking (in blue lines) and bottom ones tracking discarding non-confident points (in green lines). After 40 frames, accumulation OF errors progressively deviates the initial curve from the target ring, which corresponds to an image valley. Differences between both tracked curves can be better observed in the most right close-up image.

5.1.1 Conclusions

After applying the statistical framework presented in Chapter 4, we have improved the results in the computation of the tracheal rings of bronchoscopy images.

5.2 Cardiac Diagnose Systems. ANOVA assessment of influential factors

Changes in cardiac deformation patterns are correlated with cardiac pathologies. Deformation can be extracted from tagging Magnetic Resonance Imaging (tMRI) using Optical Flow (OF) techniques. For applications of OF in a clinical setting it is important to assess to what extent the performance of a particular OF method is stable across different clinical acquisition artifacts. This section presents a statistical validation framework, based on ANOVA, to assess the motion and appearance factors that have the largest influence on OF accuracy drop. In order to validate the application of the framework, we use a database of simulated tMRI data including the most common artifacts of MRI and test three different OF methods, including HARP.

The OF method best suited for a clinical application should be the one presenting the most stable performance across the artifacts arising in that particular clinical setting. In the context of cardiac deformation tracking, clinical settings prone to affect OF performance include, among others, variability in image acquisition conditions, radiological noise distorting image appearance and distorted motion patterns due to cardiac pathologies.

As we have seen in previous chapters, ANOVA's [89] are powerful statistical tools for detecting differences in performance across methodologies, as well as the impact of different factors or assumptions. Thus, we propose to use Analysis of Variance (ANOVA) to compare the performance of multiple OF methods and explore the impact of specific clinical conditions. In particular, we propose to use a 2-way ANOVA with factors given by the OF method (denoted OF) and the clinical source of error (denoted CSE) whose influence on OF we want to assess. The ANOVA individuals should be defined as a random sampling of consecutive frames taken from a representative set of sequences. The quantitative ANOVA variable should be a measure of OF performance computed for each of the sampled frames. Such a measure could be the pixel-based OF error summarized for the whole frame or the error in a clinical functional score calculated from OF motion (such as strain or rotation).

The desired result of the 2-ANOVA test would be a significance in the methods factor, possibly a significance across CSE and, most important, no significant interaction. In case of significant interaction ($p - val < \alpha$), a 1-way ANOVA with the combined OF-CSE factor should be used to detect the sources of bias. Otherwise, the significance of each ANOVA factor can be correctly interpreted using its associated p-value. In case of significant differences ($p - val < \alpha$), we can compare group factors using a multiple comparison test with Tukey correction [90] to detect those groups that are significantly worse. In this paper we have considered 3 different types of CSE:

- **CSE1: Acquisition impact**. The typical tMRI acquisition can produce either two sequences with complementary stripes or a single sequence combining both magnetic fields into a single grid pattern. The two patterns define the CSE groups.
- **CSE2: Radiological noise impact.** The influence of the radiological noise should be assessed by considering sequences with decreasing SNR. The different

SNR_S define the ANOVA groups of the CSE factor.

- **CSE3: Motion impact.** Finally, several kinds of pathologies should be considered in order to assess if OF methods are biased due to regularity assumptions or a priori models of motion. The different motion patterns define CSE factor groups.

As a first step towards a full validation of CSE influence using clinical data, we use a database [71] simulating the above conditions using the model of cardiac deformation and SPAMM acquisition (for more information see Chapter 2).

In this study, we choose motion estimation errors given by OF End-point-Error (EE) [48] to define the ANOVA variable. Given that EE is computed for each pixel, the ANOVA variable is the EE average: $\mu(EE) := \frac{1}{N}\sum EE_i$, with EE_i the error for each pixel and N the number of pixels. In order to account for non-normality in the data, $\mu(EE)$ was transformed to the logarithmic scale [89]. ANOVA tests were performed at a significance level $\alpha = 0.05$.

Concerning ANOVA individuals and groups, we defined them using the dataset described in Chapter 2. The CSE factor groups are given as:

- **CSE1.** We used the sequences without noise and decay (SNR_{100}) and with the full 2D motion grouped according to their tag pattern, which denoted as *grid* and *striped*. The ANOVA individuals were taken from a random sampling of 7 frames of sequences at basal, mid and apical levels. This ANOVA should assess the impact of the grid pattern under the best possible setting and it selects the pattern for the remaining experiments.

- **CSE2.** The impact of radiological noise was assessed by taking sequences without noise and without decay, denoted by SNR_{100}, and with decay and the constant Rician noise added, denoted by $SNR_{25} - D$. As before, the full 2D motion sequences with grid pattern at basal, mid and apical levels randomly sampled define the ANOVA individuals.

- **CSE3.** Finally, the impact of motion bias in OF assumptions is checked by considering 2D motion, noted by $2DF$, and its decoupling into rotation, denoted R, and contraction, denoted C, as CSE groups. Sequences were considered with Rician noise and decay to account for conditions as realistic as possible. This ANOVA should detect the impact of regularity assumptions in OF computation.

The OF factor groups are three methods working on the frequency domain and with different regularity assumptions for a fair assessment of CSE3:

- **Full HPF (HPF).** Implementation of the algorithm described by Garcia et al. in [39]. The data term is computed using the phase images of each tagging pattern and is combined with the smoothness term using variable weights given by the amplitudes of the Gabor filter responses.

- **Constant HPF (HPF_C).** Adaptation of [39] with constant weights set to 0.5.

- **HARP ($HARP$).** In-house implementation of the algorithm described by Osman et al. in [64].

CSE1			CSE2			CSE3		
$p-OF$	$p-CSE$	$p-int$	$p-OF$	$p-CSE$	$p-int$	$p-OF$	$p-CSE$	$p-int$
$\ll 10^{-16}$	0.239	0.657	$\ll 10^{-16}$	0.058	0.251	$\ll 10^{-16}$	0.852	0.874

Table 5.2
ANOVA RESULTS.

In order to avoid introducing a bias in the results, we computed harmonic phase images for all of the input images, as described in [64]. These images were then used as input for all OF methods.

Table 5.2 reports the 2-ANOVA p-values for the CSE experiments: p-OF for the OF factors, p-CSE for CSE factors, and p-int for interaction. For all experiments, there is no evidence of significant interaction ($p-int > 0.05$), but there are significant differences in OF performance ($p - OF \ll 10^{-16}$). It follows that, OF performance ranking is independent of the considered CSE conditions and the most suitable OF method can be selected. Concerning the CSE factor there are no significant differences ($p - CSE > 0.05$), so that all OF methods are robust against the clinical settings considered. However, it is worth noticing that the presence of noise causes p-values to drop so a further decrease in SNR could affect OF performance.

In order to further explore group differences and, in the particular case of OF significant differences, discard the worse methods, we have applied the pairwise comparison with Tukey correction shown in Figure 5.3. For each factor, Figure 5.3 shows group mean differences represented as horizontal lines centred at the mean (in logarithmic scale) and vertically distributed according to the factor group. Group differences being in logarithmic scale, the more negative mean values are, the smaller the OF error is. We observe that, for all CSE conditions, the best OF method is HPF and the worst one $HARP$. Regarding the impact of CSE conditions, although there is not enough evidence of differences, plots reveal some interesting tendencies. First, we observed that considering two sequences with stripe lines has smaller error in OF computations. Second, OF methods performance is better without noise and decay, as expected. Finally, there is no difference across different motions, so that OF motion assumptions do not bias computations.

5.2.1 Conclusions

We presented a validation framework that uses ANOVA to detect significant differences in OF performance according to different clinical factors prone to have large influence in OF accuracy. Our framework has been applied to quantitatively test the performance of three OF methods working on the frequency domain ($HARP$, HPF and HPF_C).

On the one hand, the presented experiments show that a method (HPF) that applies the regularity term only at areas where phase is not reliable performs better than the one using a global regularity constraint (HPF_C). Experiments also show the need for the regularity term to reduce $HARP$ sensitivity to noise.

On the other hand, experiments show that there is no bias due to CSE. First of

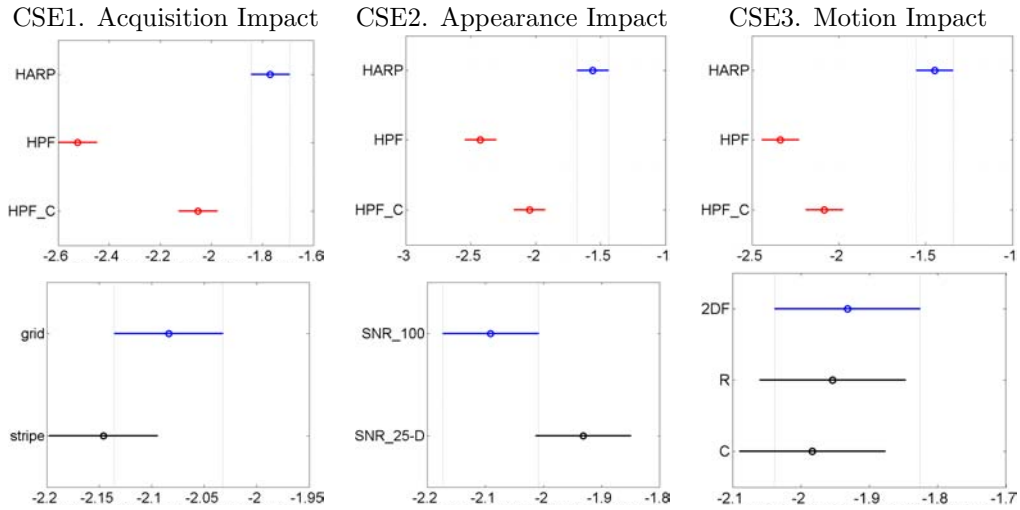


Figure 5.3: Pairwise comparison with Tukey correction. Results on the EE group mean shown in logarithmic scale in the horizontal axis.

all, using as input image stripes or a grid pattern does not affect OF performance significantly. Regarding the $SNR_{25} - D$ versus SNR_{100} sequences, despite there are no significant differences, we observe that OF performance is better for clean sequences. Finally, motion assumptions do not bias computations. Summarizing, the chosen OF methods are robust against CSE artifacts.

Chapter 6

Final Remarks and Future Work

Final remarks

For a good quality and reliability of the results of applications that use optical flow as an input, it is of prime importance the definition of a framework that is able to assess the optical flow performance. This thesis analyzes the error sources of optical flow techniques as well as the capabilities of confidence measures to bound such errors. In addition it defines qualitative scores that assess the performance of the CM for error bounding. Those qualitative scores are the previous steps to the final goal, which is the definition of a robust framework able to asses the best pair OF-CM that is able bound the optical flow error. Finally, it shows the applicability of the presented framework in medical imaging.

- **Optical flow error sources.** Some of the errors produced by the optical flow can be controlled or improved by the framework definition, but other errors are inherent to the optical flow computation, and thus, we can not get rid of them. Consequently it is of prime importance to be able to assess the reliability of the computed flow fields.
- **Confidence Measure Capabilites.** State of the art confidence measures are able to asses different types of optical flow error sources, thus we need to analyze each confidence measure performance against an specific optical flow method in order to determine its real performance for error bounding.
- **Descriptive plots.** Descriptive plots are an important tool that should be able to assess the confidence measure capabilities for error bounding. That is, they should determine for each CM value which is the risk, or the percentage of points that have large errors. Thus, the presented SDP plot is the best suited tool to assess CM bounding capabilities.
- **Scores for descriptive plots.** The presented scores assess the profile of a descriptive plot, they provide a first insight in the evaluation of the capabilities of a CM to properly bound the OF error. Such scores are a required tool to the definition of the presented statistical framework to properly perform the ANOVA tests.

- **Statistical framework definition.** This work constitutes a first new effort in the use of statistical tools for the evaluation of the capabilities of CM for predicting OF error in decision support systems. Unlike machine learning approaches that require a high and large training data set capturing the variability of frames, our methodology can be applied to a reduced number of training samples and, still, provide results generalizable enough. This is a main advantage of using parametric statistical models, which are able to detect significant differences from a very reduced set of samples. In this context, our methodology is a powerful tool to perform a first exploratory analysis to discard unfeasible parameter settings in methods or algorithms ill-posed under specific application settings.

- **Framework applicability.** As any other statistical approach, our methodology can be applied as far as the training sequences are representative enough of the dynamical appearance and resolution of the application frames. This might require either a data set capturing the variability of frames appearance or controlled acquisition conditions. The latter are frequently achieved in protocols used in medical imaging or the vision systems of industrial applications.

  The potential of the presented statistical analysis for medical application has been shown in a preliminary study for the design of an optical flow method better suited for the computation of cardiac clinical scores of diagnostic value from Magnetic Resonance sequences. In addition we applied it to improve the results on the navigation inside the bronchial tree on bronchoscopy by means of improving the bronchial rings computation.

Future work

It is worth noticing that the presented framework is an statistical tool flexible enough to accommodate different applications other than optical flow, that also require to assess errors in computations. To do so, the only requirement is to have a quality that might correlate to errors. In such case, our methodology could be used to analyze its actual capabilities to bound the error and select the most appropriate threshold.

In this context, our methodology could be applied, for instance, to sparse feature tracking for camera pose estimation. Camera point of view given by the fundamental matrix is estimated from correspondent points, estimated using SIFT or SURF, in two consecutive frames. A main concern is the sensitivity to noise, which introduces outliers in matches. The most accepted method to remove outliers from a set of data is RANdom SAmple Consensus [88] (RANSAC). The basic idea is to compute the fundamental matrix from randomly selected matches and use it to project epipolar lines from one frame onto the next one. The Euclidean distance between matched points and their corresponding epipolar lines is used to detect inliers and, thus, it is a measure of the confidence in matched points. On-going work is the application of the presented framework to assess the relation of such distance to the fundamental matrix error and set a threshold ensuring enough inliers in the context of tracheal ring tracking.

In addition, the presented methodology could be used in a constructive way, either

to design a novel confidence measure which outperforms the ones considered here, or to better compare and understand the failures and successes of individual optical flow methods to design approaches that are more reliable for a decision support system.

In the following points we present some future work related with the improvement and other possibilities on the definition of the presented framework.

- **Improvement and definition of CM.** Most of existing CM are designed to assess OF model assumptions (motion and data term). Sequences where model assumptions are not fulfilled can not be used to predict error bound for that particular OF-CM pair. However, this does not invalidate our confidence framework since it could be used to statistically assess the capabilities of new designed CM for detecting any failure to satisfy data term assumptions. Fulfilment of model assumptions could be detected from appearance and temporal features [54] in order to define new CM according to such characteristics [91]. Or alternatively, sequences could be corrected to satisfy some of the general constrains (such as illumination constancy). Then, by applying the presented framework, we could assess the benefits of the correction, and also, to improve the error bound curves.

- **Impact of statistical model assumptions in prediction capabilities.** The main assumptions of our statistical analysis is that data should be unimodal and normally distributed. Concerning the first assumption, unimodality is violated for sequences containing very abrupt changes along the sequence, converting it like two different sequences, the curves will be grouped in two sets, so that the variability will be too large. In the second case, the population confidence interval used to compute $\mathcal{SDP}$ assumes that data are normally distributed. This holds for most cases. An alternative could be to use empirical percentiles to get more stable bounds. However, empirical descriptive statistics require a larger sample set ($n > 30$) in order to compute high percentiles comparable to the confidence $1 - \alpha$.

- **Alternative statistical tools.** Given the descriptive nature of the different strategies proposed, several complementary statistical indexes could be used. In this sense, in order to describe the variability of the EE values given a OF-CM, a Lorentz curve provides an interesting summary. Thus, the area under this curve (Gini's index) would be a measure of the inequalities between EE values (actually these tools are commonly used in the macro-economic field to describe income inequalities in a given country). In the same way, a global measure of the relationship between EE values and CM could be obtained by means of correlation, either non-parametric or parametric.

- **Dependency across EE_{max}.** Since the approach used depends on the choice of EE_{max}, an initial evaluation of the relationship between CM values and risk could be performed. The ROC curve is a representation of the false positive rate obtained for different false negative rates. The area under its curve summarizes the accuracy of the CM to predict large errors for a given EE_{max}. Thus, one could represent AUC(EE_{max}) vs EE_{max} to check the accuracy stability of different EE_{max} choices.

List of Publications

International Conferences

- H. Kause, P. Márquez-Valle, A. Fuster, A. Hernàndez-Sabaté, L. Florack, D. Gil, H. van Assen, "Quality Assessment of Optical Flow in Tagging MRI", *Dutch Bio-Medical Engineering Conference*, 2015. (Oral).

- P. Márquez-Valle, H. Kause, A. Fuster, A. Hernàndez-Sabaté, L. Florack, D. Gil, H. van Assen, "Factors Affecting Optical Flow Performance in Tagging Magnetic Resonance Imaging", *Medical Image Computing and Computer Assisted Intervention - Workshops*, pp. 231-238, 2014. (Oral).

- P. Márquez-Valle, D. Gil, R. Mester, A. Hernàndez-Sabaté, "Local Analysis of Confidence Measures for Optical Flow Quality Evaluation", *International Conference on Computer Vision Theory and Applications*, pp. 450-457, 2014. (Oral).

- P. Márquez-Valle, D. Gil, A. Hernàndez-Sabaté, "Evaluation of the Capabilities of Confidence Measures for Assessing Optical Flow Quality", *IEEE International Conference on Computer Vision - Workshops*, pp. 624-631, 2013. (Oral).

- P. Márquez-Valle, D. Gil, A. Hernàndez-Sabaté, Daniel Kondermann, "When Is a Confidence Measure Good Enough?", *International Conference on Computer Vision Systems*, pp. 344-353, 2013. (Poster).

- P. Márquez-Valle, D. Gil, A. Hernàndez-Sabaté, "A Complete Confidence Framework for Optical Flow", *European Conference on Computer Vision - Workshops*, pp. 124-133, 2012. (Poster).

- P. Márquez-Valle, D. Gil, A. Hernàndez-Sabaté, "Error Analysis for Lucas-Kanade Based Schemes", *International Conference on Image Analysis and Recognition*, pp. 184-191, 2012. (Oral).

- P. Márquez-Valle, D. Gil, A. Hernàndez-Sabaté, "A confidence measure for assessing optical flow accuracy in the absence of ground truth", *International Conference on Computer Vision - Workshops*, pp. 2042-2049, 2011. (Oral).

Teaching Publications

- C. Sánchez, O. Ramos, P. Márquez-Valle, E. Martí, J. Rocarías, D. Gil, "Automatic evaluation of practices in Moodle for Self Learning in Engineering", *Journal of Teaching and Science Education*, pp. to appear, 2015.
- C. Sánchez, O. Ramos, P. Márquez-Valle, E. Martí, J. Rocarías, D. Gil, "Evaluación automática de prácticas en Moodle para el aprendizaje autónomo en Ingenierías ", *International Congress on University Teaching and Innovation*, 2014. (Oral discussion).

Bibliography

[1] M. Choi, J. Park, and S. Lee, "Event classification for vehicle navigation system by regional optical flow analysis," *Machine Vision and Applications*, vol. 25, no. 3, pp. 547–559, 2014.

[2] J. Xu, H. Chen, W. Ding, C. Zhao, and J. Morris, "Robust optical flow for driver assistance," in *IVCNZ*. IEEE, 2010, pp. 1–7.

[3] N. Onkarappa and A. Sappa, "Speed and texture: An empirical study on optical-flow accuracy in adas scenarios," 2014.

[4] A. Nolan, D. Serrano, A. Hernandez, and D. Ponsa & A. Lopez, "Obstacle mapping module for quadrotors on outdoor search and rescue operations," in *IMAV*, 2013.

[5] Y. Yang, Q. Liu, R. Ji, and Y. Gao, "Dynamic 3d scene depth reconstruction via optical flow field rectification," *PLoS ONE*, vol. 7(11), 2012.

[6] Y. Diskin and V. Asari, "Dense 3d point-cloud model using optical flow for a monocular reconstruction system," *IEEE Applied Imagery Pattern Recognition Workshop: Sensing for Control and Augmentation*, pp. 1–6, 2013.

[7] P. Sundberg, T. Brox, M. Maire, P. Arbelaez, and J. Malik, "Occlusion boundary detection and figure/ground assignment from optical flow," in *IEEE CVPR*, 2011.

[8] C. Ballester, L. Garrido, V. Lazcano, and V. Caselles, "A TV-L1 optical flow method with occlusion detection," in *DAGM/OAGM Symposium*. 2012, vol. 7476 of *Lecture Notes in Computer Science*, pp. 31–40, Springer.

[9] W. Li, X. Wu, K. Matsumoto, and H. Zhao, "Foreground detection based on optical flow and background subtract," in *International Conference on Communications, Circuits and Systems*. 2010, pp. 359–362, IEEE.

[10] K. Suganya-Devi, N. Malmurugan, and R. Sivakumar, "Efficient foreground extraction based on optical flow and smed for road traffic analysis," in *International Journal of Cyber-Security and Digital Forensics*, 2012, vol. 1(3), pp. 177–182.

[11] D. Gil, O. Rodriguez-Leor, P. Radeva, and J. Mauri, "Myocardial perfusion characterization from contrast angiography spectral distribution," *Medical Imaging, IEEE Transactions on*, vol. 27, no. 5, pp. 641–649, 2008.

[12] R. Eberhardt, N. Kahn, D. Gompelmann, M. Schumann, C. Heussel, and F. Herth, "Lungpoint-a new approach to peripheral lesions," *Journal of Thoracic Oncology*, vol. 5, no. 10, pp. 1559–1563, 2010.

[13] A. Hernandez-Sabate, D. Gil, J. Garcia-Barnes, and E. Marti, "Image-based cardiac phase retrieval in intravascular ultrasound sequences," *Ultrasonics, Ferroelectrics, and Frequency Control, IEEE Transactions on*, vol. 58, no. 1, pp. 60–72, 2011.

[14] X. Luó, M. Feuerstein, D. Deguchi, T. Kitasaka, H. Takabatake, and K. Mori, "Development and comparison of new hybrid motion tracking for bronchoscopic navigation.," *Medical Image Analysis*, vol. 16, no. 3, pp. 577–596, 2012.

[15] F. Asano, N. Shinagawa, T. Ishida, J. Shindoh, M. Anzai, A. Tsuzuku, S. Oizumi, and S. Morita, "Virtual bronchoscopic navigation combined with ultrathin bronchoscopy. a randomized clinical trial," *American journal of respiratory and critical care medicine*, vol. 188, no. 3, pp. 327–333, 2013.

[16] R. Duits, B. Janssen, A. Becciu, and H. Van Assen, "A variational approach to cardiac motion estimation based on covariant derivatives and multi-scale helmholtz decomposition," *Q. Appl. Math.*, vol. 71, no. 1, pp. 1–36, 2013.

[17] J. Garcia-Barnes, D. Gil, L. Badiella, A. Hernandez-Sabate, F. Carreras, S. Pujades, and E. Martí, "A normalized framework for the design of feature spaces assessing the left ventricular function," *Medical Imaging, IEEE Transactions on*, vol. 29, no. 3, pp. 733–745, 2010.

[18] A. Hernandez, P. Radeva, A. Tovar, and D. Gil, "Vessel structures alignment by spectral analysis of ivus sequences," *Proc. of CVII, MICCAI Workshop*, 2006.

[19] M. Götte, T. Germans, I. Rüssel, J. Zwanenburg, J. Marcus, A. van Rossum, and D. van Veldhuisen, "Myocardial strain and torsion quantified by cardiovascular magnetic resonance tissue taggingstudies in normal and impaired left ventricular function," *Journal of the American College of Cardiology*, vol. 48, no. 10, pp. 2002–2011, 2006.

[20] F. Carreras, J. Garcia-Barnes, D. Gil, S. Pujadas, C. Li, R. Suarez-Arias, R. Leta, X. Alomar, M. Ballester, and G. Pons-Llado, "Left ventricular torsion and longitudinal shortening: two fundamental components of myocardial mechanics assessed by tagged cine-mri in normal subjects," *The international journal of cardiovascular imaging*, vol. 28, no. 2, pp. 273–284, 2012.

[21] H. Colt and S. Murgu, *Bronchoscopy and central airway disorders*, Elsevier, 2012.

[22] S. Murgu and H. Colt, "Morphometric bronchoscopy in adults with central airway obstruction: Case illustrations and review of the literature," *The Laryngoscope*, vol. 119, no. 7, pp. 1318–1324, 2009.

[23] S. Murgu and H. Colt, "Subjective assessment using still bronchoscopic images misclassifies airway narrowing in laryngotracheal stenosis," *Interactive cardiovascular and thoracic surgery*, vol. 16, no. 5, pp. 655–660, 2013.

[24] A. Begnaud, J. Connett, E. Harwood, M. Jantz, and H. Mehta, "Measuring central airway obstruction: What do bronchoscopists do?," *Annals of the American Thoracic Society*, 2014.

[25] C. Sánchez, J. Bernal, F. Sánchez, M. Diez, A. Rosell, and D. Gil, "Toward online quantification of tracheal stenosis from videobronchoscopy," *International journal of computer assisted radiology and surgery*, pp. 1–11, 2015.

[26] B. Ristic, S. Arulampalam, and N. Gordon, *Beyond the Kalman Filter: Particle Filters for Tracking Applications*, Artech House, 2004.

[27] L. Morency, J. Whitehill, and J. Movellan, "Monocular head pose estimation using generalized adaptive view-based appearance model," *Image and Vision Computing*, vol. 28, no. 5, pp. 754–761, 2010.

[28] J. J. Gibson, *The Perception of the Visual World*, Riverside Press, Cambridge, 1950.

[29] J. J. Gibson, *The Senses Considered as Perceptual Systems*, Houghton-Mifflin, Boston, 1966.

[30] D. Warren and E. Strelow, *Electronic Spatial Sensing for the Blind: Contributions from Perception, Rehabilitation, and Computer Vision*, Springer Netherlands, 1985.

[31] J. L. Barron, D. J. Fleet, and S. S. Beauchemin, "Performance of optical flow techniques," *IJCV*, vol. 12, no. 1, pp. 43–77, 1994.

[32] S. Beauchemin and J. Barron, "The computation of optical flow," *ACM Comput. Surv.*, vol. 27, no. 3, pp. 433–466, 1995.

[33] B. Lucas and T. Kanade, "An iterative image registration technique with an application to stereovision," in *DARPA Image Understanding Workshop*, 1981, pp. 121–130.

[34] D. J. Fleet and A. D. Jepson, "Computation of component image velocity from local phase information," *IJCV*, vol. 5, no. 1, pp. 77–104, 1990.

[35] B. Horn and B. Schunck, "Determining optical flow," *AI*, vol. 17, pp. 185–203, 1981.

[36] H. H. Nagel and W. Enkelmann, "An investigation of smoothness constraints for the estimation of displacementvectorfields from image sequences," *PAMI*, vol. 8, pp. 565–593, 1986.

[37] J. Bigün, G. H. Granlund, and J. Wiklund, "Multidimensional orientation estimation with applications to texture analysisand optical flow," *PAMI*, vol. 13, no. 8, pp. 775–790, 1991.

[38] A. Bruhn, J. Weickert, and C. Schnörr, "Lucas/Kanade meets Horn/Schunck: Combining local and global opticflow methods," *IJCV*, vol. 61, no. 2, pp. 221–231, 2005.

[39] J. Garcia-Barnes, D. Gil, S. Pujades, and F. Carreras, "Variational framework for assessment of the left ventricle motion," *Math. Mod. of Nat. Phen.*, vol. 3, no. 6, pp. 76–100, 2008.

[40] D. Sun, S. Roth, and M. J. Black, "Secrets of optical flow estimation and their principles," in *CVPR*, 2010, pp. 2432–2439.

[41] J. Weickert and C. Schnörr, "Variational optic flow computation with spatio-temporal smoothness constraint," *JMIV*, vol. 14, no. 3, pp. 245:255, 2001.

[42] S. Roth and M.J. Black, "On the spatial statistics of optical flow," *IJCV*, vol. 74, no. 1, pp. 33–50, 2007.

[43] C. Zach, T. Pock, and H. Bischof, "A duality based approach for real-time TV-L1 optical flow," *PR (Proc. DAGM)*, pp. 214–223, 2007.

[44] M. Werlberger, T. Pock, and H. Bischof, "Motion estimation with non-local total variation regularization," in *CVPR*, 2010, pp. 2464–2471.

[45] H. Zimmer, A. Bruhn, and J. Weickert, "Optical flow in harmony," *IJCV*, vol. 93, pp. 368–388, 2011.

[46] D. Kondermann, S. Abraham, G. Brostow, et al., "On performance analysis of optical flow algorithms," *Outdoor and Large-Scale Real-World Scene Analysis*, pp. 329–355, 2012.

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

[48] S. Baker, D. Scharstein, J Lewis, and et al., "A database and evaluation methodology for optical flow," *IJCV*, vol. 92, no. 1, pp. 1–31, 2011.

[49] A. Bruhn and J. Weickert, "A confidence measure for variational optic flow methods," in *Geometric Properties for Incomplete Data*, 2006, pp. 283–298.

[50] A. Singh, "An estimation-theoretic framework for discontinuous flow fields," in *ICCV*, 1990, pp. 168–177.

[51] J. Shi and C. Tomasi, "Good features to track," pp. 593–600, 1994.

[52] C. Kondermann, R. Mester, and C. Garbe, "A statistical confidence measure for optical flows," in *ECCV*, 2008, pp. 290–301.

[53] J. Kybic and C. Nieuwenhuis, "Bootstrap optical flow confidence and uncertainty measure," *CVIU*, pp. 1449–1462, 2011.

[54] O. Mac Aodha, A. Humayun, M. Pollefeys, and G. J. Brostow, "Learning a confidence measure for optical flow," *IEEE PAMI*, vol. 35, pp. 1107–1120, 2013.

[55] H. Liu, T. Hong, M. Herman, and R. Chellappa, "A general motion model and spatio-temporal filters for computing opticalflow," *International Journal of Computer Vision*, vol. 22, no. 2, pp. 141–172, 1997.

[56] S. Fazekas and D. Chetverikov, "Analysis and performance evaluation of optical flow features for dynamictexture recognition," *Signal Processing: Image Communication*, vol. 22, no. 7, pp. 680–691, 2007.

[57] M. Everingham, S. M. Ali Eslami, L. Van Gool, C. K.I. Williams, J. Winn, and A. Zisserman, "Assessing the significance of performance differences on the pascal vocchallenges via bootstrapping," Tech. Rep., 2013.

[58] P. Anandan, "A computational framework and an algorithm for the measurement of visualmotion," *IJCV*, vol. 2, pp. 283–310, 1989.

[59] B. F. Buxton and H. Buxton, "Computation of optic flow from the motion of edges in the image sequences," *Image and Vision Computing*, vol. 2, pp. 59–75, 1984.

[60] A. M. Waxman, J. Wu, and F. Bergholm, "Convected activation profiles and the measurement of visual motion," *Proc. IEEE CVPR*, pp. 717–723, 1988.

[61] J. Wills, S. Agarwal, and S. Belongie, "A feature-based approach for dense segmentation and estimation of largedisparitymotion," *IJCV*, vol. 68, no. 2, pp. 125–143, 2006.

[62] M. Felsberg, "Optical flow estimation from monogenic phase," *IWCM*, vol. LNCS 3417, pp. 1–13, 2004.

[63] J. Garcia-Barnes, *Statistical Models of the Architecture and Function of the Left Ventricle*, Ph.D. thesis, Autonomous University of Barcelona, Bellaterra, Spain, 2009.

[64] N. F. Osman, W. S. Kerwin, E. R. McVeigh, , and J. L. Prince, "Cardiac motion tracking using using CINE harmonic phase (HARP) magneticresonanceimaging," *Magnetic Resonance in Medicine*, vol. 42, pp. 1048–1060, 1999.

[65] N. Osman, E. McVeigh, and J. Prince, "Imaging heart motion using harmonic phase mri.," *IEEE Trans. Med. Imaging*, vol. 19, no. 3, pp. 186–202, 2000.

[66] A. Wedel, T. Pock, C. Zach, H. Bischof, and D. Cremers, "An improved algorithm for TV-L1 optical flow," *Statistical and Geometrical Approaches to Visual Motion Analysis: InternationalDagstuhlSeminar*, pp. 23–45, 2009.

[67] M. Drulea and S. Nedevschi, "Motion estimation using the correlation transform," *IEEE Transactions on Image Processing*, vol. 22, no. 8, pp. 3260–3270, 2013.

[68] D. Butler, J. Wulff, G. Stanley, and M. Black, "A naturalistic open source movie for optical flow evaluation," in *ECCV*. 2012, 7577, pp. 611–625, Springer-Verlag.

[69] O. Mac Aodha, G. Brostow, and M. Pollefeys, "Segmenting video into classes of algorithm-suitability," in *CVPR*, 2010.

[70] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[71] P. Márquez-Valle, H. Kause, A. Fuster, A. Hernàndez-Sabaté, L. Florack, D. Gil, and H. van Assen, "Factors affecting optical flow performance in tagging magnetic resonance imaging," in *MICCAI - Workshops*, 2014, pp. 231–238.

[72] T. Arts, W Hunter, A. Douglas, and et al., "Description of the deformation of the left ventricle by a kinematic model.," *Journal Biomechanics*, vol. 25, no. 10, pp. 1119–1127, 1992.

[73] E. Waks, J. Prince, and A. Douglas, "Cardiac motion simulator for tagged MRI," in *MMBIA-Workshops*, 1996, pp. 182–191.

[74] M. Gutberlet, K. Schwinge, P. Freyhardt, and et al., "Influence of high magnetic field strengths and parallel acquisition strategies on image quality in cardiac 2D CINE magnetic resonance imaging," *Eur Radiol*, vol. 15, no. 8, pp. 1586–97, 2005.

[75] H. Haußecker, H. Spies, and B. Jähne, "Tensor-based image sequence processing techniques for the study of dynamicalprocesses," in *Proc. Int. Symp. On Real-Time Imaging and Dynamic Analysis*. Citeseer, 1998, vol. 32, pp. 704–711.

[76] John W Tukey, "The philosophy of multiple comparisons," *Statistical science*, pp. 100–116, 1991.

[77] Lawrence C. Evans, *Partial Differential Equations*, American Mathematical Society, 1998.

[78] R. Burden and J. Douglas Faires, *Numerical Analysis*, Thompson, 2005.

[79] P. Márquez-Valle, D. Gil, and A. Hernàndez-Sabaté, "A complete confidence framework for optical flow," in *European Conference on Computer Vision - Workshops*, 2012, pp. 124–133.

[80] E. Parzen, "On estimation of a probability density function and mode," *The Annals of Mathematical Statistics*, vol. 33, no. 3, pp. 1065–1076, 1962.

[81] M. S. Pepe, *The Statistical Evaluation of Medical Tests for Classification and Prediction*, Oxford Statistical Sciences Series, 2003.

[82] C. Liu, *Beyond pixels: exploring new representations and applications for motionanalysis*, Ph.D. thesis, Cambridge, MA, USA, 2009.

[83] R. Fisher, *Statistical Methods and Scientific Inference*, Oliver and Boyd, 1956.

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

[85] J. Cohen, *Statistical powe analysis for the behavioral sciences*, Lawrence Erlbaum Associates, 1988.

[86] L. Brown, T. Cai, and A. DasGupta, *Interval Estimation for a Binomial Proportion*, vol. 16, Statistical Science, 2001.

[87] D. Kondermann, C.and Kondermann, B. Jähne, and C. Garbe, "An adaptive confidence measure for optical flows based on linear subspaceprojections.," in *DAGM-Symposium*, 2007, vol. 4713 of *LNCS*, pp. 132–141.

[88] M. Fischler and R. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[89] S.F. Arnold, *The theory of linear models and multivariate observations*, Wiley, 1997.

[90] L. Tukey, "Comparing individual means in the analysis of variance," *Biometrics*, vol. 5, pp. 99–114, 1949.

[91] O. Mac Aodha, G. Brostow, and M. Pollefeys, "Segmenting video into classes of algorithm-suitability," in *CVPR*, 2010.