Factors Affecting Optical Flow Performance in Tagging Magnetic Resonance Imaging

Patricia Márquez-Valle[1], Hanne Kause[2], Andrea Fuster[3,4], Aura Hernàndez-Sabaté[1], Luc Florack[3,4], Debora Gil[1], and Hans van Assen[2]

[1] Computer Vision Center, Autonomous University of Barcelona, Spain

[2] Department of Electrical Engineering
[3] Department of Biomedical Engineering
[4] Department of Mathematics and Computer Science
Eindhoven University of Technology, The Netherlands

pmarquez@cvc.uab.es

Abstract. 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 paper 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 this framework, we created a database of simulated tMRI data including the most common artifacts of MRI and test three different OF methods, including HARP.

Keywords: Optical Flow, Performance Evaluation, Synthetic Database, ANOVA, Tagging Magnetic Resonance Imaging

1 Introduction

Tagging MRI (tMRI) is an important imaging technique that enables detailed motion analysis of the cardiac left ventricle (LV) [1]. Tagging MR images can be obtained by spatially modulating the MR magnetization field [2] so that images have a characteristic stripe or grid pattern that deforms along with cardiac tissue. This makes it possible to track information about motion and deformation over time, alterations of which are known to correlate with pathology [3–5].

A well-known technique to obtain motion information from image sequences is optical flow (OF), which results in a dense motion field by minimizing an energy functional that combines a data and a smoothness term [6]. OF techniques have two main types of error sources: model assumptions and numerics. Model assumption errors arise when images do not meet the expected appearance or motion patterns, such as the brightness constancy violated by signal decay, or flow field irregularity. Numerical errors arise from the propagation of errors in the input data to errors in the output, e.g. signal decay or noise. Although it is impossible to avoid these errors, as they are inherent to any real world problem, some model assumptions, such as brightness constancy, can be modelled in the case of tMRI by using the harmonic phase of the original signal, which is constant over time [7–12]. Variable brightness OF has been studied in [13], where the signal decay was modelled by including decay terms in the OF equations. Still, resulting flow vectors can not be made error free.

Much work has been done on OF for tMRI [14–16]. However, signal decay and noise (among others) will always influence its accuracy [17]. In order to correctly interpret results it is, therefore, important to provide a quantitative estimate of the impact of the most influencing factors in OF accuracy. A qualitative comparison of four different algorithms for automatic motion analysis of the heart was carried out in [17]. The performance was assessed by a visual comparison of the box-plots as a function of SNR. From that paper, one can discern that the most stable method is the one working in the frequency domain. Still, it is worth noting that the extent to which the methods are stable in the presence of artifacts needs to be quantified.

In order to do so, as well as to be able to infer the OF accuracy range in the absence of ground truth (real images), we present in this paper a validation framework using Analysis of Variance (ANOVA [18]). This tool usually detects differences in performance, and evaluates the impact of different factors or assumptions across methodologies used for new diagnostic scores [19]. In this paper we do a step forward in the use of this statistical tool. We assess the performance of different algorithms against several factors using a 2-way ANOVA. In particular, we use ANOVA to evaluate the impact of clinical acquisition conditions (noise, signal decay, tagging acquisition), motion patterns and the influence on the results of the regularization built into some OF methods. The performance has been evaluated on three OF algorithms, two Harmonic Phase Flow (HPF) [11] implementations and HARP [7]. Our framework, presented in Section 2, is applied to a database of synthetic tagged MR images (subsection 2.1). This database contains realistic tagging images, which are either line tagged or grid tagged and with several motion patterns based on a cardiac motion simulator by Arts et al. [20]. Section 3 presents the experiments and section 4 the conclusions.

2 ANOVA assessment of influential factors

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.

We propose to use Analysis of Variance (ANOVA) to compare the performance of multiple OF methods and explore the impact of specific clinical conditions. ANOVA's [18] are powerful statistical tools 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, different factors and methods are determined. The quantitative score taken from a sampling (individuals) of the clinical data is grouped according to such factors and differences among a quantitative response group mean are computed. ANOVA provides a statistical way to assess if such differences are significant with a given confidence level α.

In order to assess the impact of a given clinical setting on the performance of several OF methods, 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 [21] 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 have simulated the above conditions using the model of cardiac deformation and SPAMM acquisition described in what follows:

2.1 Synthetic image database construction

To test the framework described above, it is necessary to have images with a ground truth motion field. For this purpose we defined a database of synthetic MR images (Fig. 1), by making use of the cardiac motion simulator by Arts and Waks [20, 22]. The heart motion was modelled using their 13-parameter model which includes radially-dependent compression, torsion, ellipticallization, shear, rotation and translation.

The initial shape of the left ventricle was modelled as a prolate sphere. A total number of 6.300 material points was sampled on the model and the deformation of the initial shape is computed using a time-dependent parametric model. We adapted the full 3D model by eliminating longitudinal motion to prevent out-of-plane motion. Different image datasets were created restricting motion to either rotation around the z-axis or radially-dependent contraction.

Although the model allows for slices with any orientation, we created datasets consisting of five short axis slices sampled across the prolate sphere. Every slice had 50×50 isotropic pixels and started with the longitudinal axis in the center of the image. For construction of the images from a set of points of transformed material coordinates with intensities, a linear interpolation approach was used. The cardiac cycle was split into 16 frames, but this can be further sub-sampled to account for low temporal resolution.

The datasets contained three sinusoidal SPAMM (spatial modulation of magnetization) [2] tagging sequences, two stripe tagging sequences (horizontal and vertical) and a grid sequence, that were modelled with signal decay according to [22].

The spatial period of the tagging patterns was set to 6.6 pixels. Rician noise was 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 [23]. Two data variants have been generated: images with noise and decay and clean data without noise and decay.

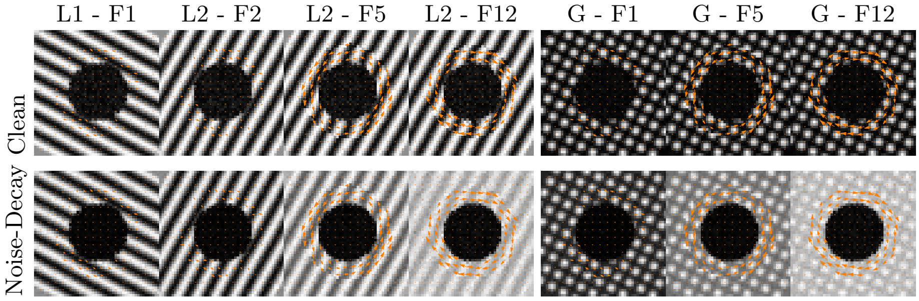


Fig. 1. Line (L1 and L2) and grid (G) tagging images (frames 1, 5 and 12) from the 3rd slice of the set with 2D cardiac motion. Clean data is shown above and data with noise and decay below. Arrows illustrate a sample of the ground truth.

3 Experiments

In this study, we choose motion estimation errors given by OF End-point-Error[5] (EE) [24] 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 [18]. ANOVA tests were performed at a significance level $\alpha = 0.05$.

Concerning ANOVA individuals and groups, we defined them using the dataset described in section 2.1. 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.

[5] $EE := \sqrt{(U-u)^2 + (V-v)^2}$ for (U, V) the ground truth flow field, and (u, v) the flow field computed using a given OF to be tested.

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 [11]. 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 [11] with constant weights set to 0.5.
- **HARP (*$HARP$*).** In-house implementation of the algorithm described by Osman et al. in [7].

In order to avoid introducing a bias in the results, we computed harmonic phase images for all of the input images, as described in [7]. These images were then used as input for all OF methods.

Table 1 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.

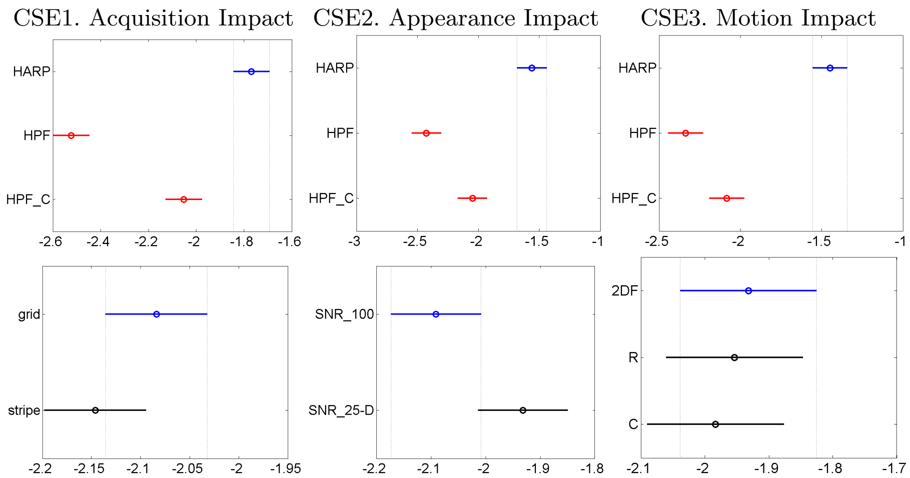
In order to further explore group differences and, in the particular case of OF significant differences, discard the worse methods, we have applied the pairwise comparison with Tukey correction shown in Figure 2. For each factor, Figure 2 shows group mean differences represented as horizontal lines centred 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.

4 Conclusions

We presented a validation framework that uses ANOVA to detect significant differences in OF performance according to different clinical factors prone to have

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 1. ANOVA results.



CSE1. Acquisition Impact CSE2. Appearance Impact CSE3. Motion Impact

Fig. 2. Pairwise comparison with Tukey correction. Results on the EE group mean shown in logarithmic scale in the horizontal axis.

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 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.

This preliminary study encourages the use of the presented framework to explore OF performance in new settings. In the future we aim apply it to the clinical sequences that were made available in the 2011 STACOM motion tracking challenge to assess the impact of tMRI features on the computation of scores of potential diagnostic value. This will also make enable comparison with other methods previously tested on this dataset.

Acknowledgements Work supported by Spanish project TIN2012-33116. First author supported by FPI-MICINN BES-2010-031102 program. Second author is supported by the Dutch Technology Foundation STW, which is part of the Netherlands Organisation for Scientific Research (NWO), and which is partly funded by the Ministry of Economic Affairs. The authors would like to thank Marijke Dermois for support with the artificial tagging MR images.

References

1. Zerhouni, E., Parish, D., Rogers, W., et al.: Human heart: Tagging with MR imaging-a method for noninvasive assessment of myocardial motion. Radiology **169**(1) (1988) 59–63
2. Axel, L., Dougherty, L.: MR imaging of motion with spatial modulation of magnetization. Radiology **171**(3) (1989) 841-845
3. Mirsky, I., Pfeffer, J., Pfeffer, M., Braunwald, E.: The contractile state as the major determinant in the evolution of left ventricular dysfunction in the spontaneously hypertensive rat. **53** (1983) 767–778
4. Götte, M., van Rossum, A., Twisk, J., et al.: Quantification of regional contractile function after infarction: Strain analysis superior to wall thickening analysis in discriminating infarct from remote myocardium. JACC **37** (2001) 808–817
5. Delhaas, T., Kotte, J., van der Toorn, A., et al.: Increase in left ventricular torsion-to-shortening ratio in children with valvular aorta stenosis. **51** (2004) 135–139
6. Horn, B., Schunck, B.: Determining optical flow. AI **17** (1981) 185–203
7. Osman, N., Kerwin, W., McVeigh, E., Prince, J.: Cardiac motion tracking using CINE HARP magnetic resonance imaging. **42**(6) (1999) 1048–1060
8. Prince, J., McVeigh, E.: Motion estimation from tagged MR image sequences. **11**(2) (1992) 238–249
9. Florack, L., van Assen, H.: A new methodology for multiscale myocardial deformation and strain analysis based on tagging MRI. (2010)
10. Xavier, M., Lalande, A., Walker, P., et al.: An adapted optical flow algorithm for robust quantification of cardiac wall motion from standard cine-MR examinations. Inf. Tech. in Biomed. **16**(5) (2012) 859–868
11. Garcia-Barnes, J., Gil, D., Pujades, S., Carreras, F.: Variational framework for assessment of the left ventricle motion. Math. Mod. of Nat. Phen. **3**(6) (2008) 76–100
12. Becciu, A., van Assen, H., Florack, L., et al.: A multi-scale feature based optic flow method for 3D cardiac motion estimation. In: SSVM. (2009) 588–599
13. Gupta, S., Prince, J.: On variable brightness optical flow for tagged MRI. (1995) 323–334
14. Carranza-Herrezuelo, N., Bajo, A., Sroubek, F., et al.: Motion estimation of tagged cardiac magnetic resonance images using variational techniques. Comp. Med. Im. and Graph. **34**(6) (2010) 514–522
15. Alessandrini, M., Basarab, A., Liebgott, H., Bernard, O.: Myocardial motion estimation from medical images using the monogenic signal. Image Processing, IEEE Transactions on **22**(3) (2013) 1084–1095
16. Arts, T., Prinzen, F., Delhaas, T., et al.: Mapping displacement and deformation of the heart with local sine-wave modeling. **29**(5) (2010)

17. Smal, I., Carranza-Herrezuelo, N., Klein, S., et al.: Reversible jump MCMC methods for fully automatic motion analysis in tagged MRI. Medical Image Analysis **16**(1) (2012) 301–324
18. Arnold, S.: The theory of linear models and multivariate observations. Wiley (1997)
19. Jap, B., Lal, S., P.Fischer, Bekiaris, E.: Using eeg spectral components to assess algorithms for detecting fatigue. Expert Systems with Applications **36**(2) (2009) 2352–2359
20. Arts, T., Hunter, W., Douglas, A., et al.: Description of the deformation of the left ventricle by a kinematic model. Journal Biomechanics **25**(10) (1992) 1119–1127
21. Tukey, L.: Comparing individual means in the analysis of variance. Biometrics **5** (1949) 99–114
22. Waks, E., Prince, J., Douglas, A.: Cardiac motion simulator for tagged MRI. In: MMBIA-Workshops. (1996) 182–191
23. Gutberlet, M., Schwinge, K., Freyhardt, P., et al.: Influence of high magnetic field strengths and parallel acquisition strategies on image quality in cardiac 2D CINE magnetic resonance imaging. Eur Radiol **15**(8) (2005) 1586–97
24. Baker, S., Scharstein, D., Lewis, J., et al.: A database and evaluation methodology for optical flow. IJCV **92**(1) (2011) 1–31