Understanding the High Accuracy Optical Flow Estimation based on a Theory for Warping*

Anders Nylander

October 31, 2016

Abstract

As an assignment, I read and summarize a paper related to optical flow estimation. The authors use an energy functional for computing optical flow that combines three assumptions: brightness and gradient constancy assumptions, and a discontinuity preserving spatio-temporal smoothness constraint. Using this model, they derive a high accuracy and fast method for optical flow and motion estimation by postponing the linearisation of the problem from the model to the numerical approximation. By using a coarse-to-fine warping technique, they prove that warping methods are theoretically justified for numerical approximation. This proposed method is shown to produce excellent results, is highly insensitive to parameter changes, and is very robust against noise.

1 Introduction

In the course "Computer Vision in 3D" teached at Halmstad University I was given a task by the course examiner to write a report on a scientific paper related to optical flow, with the requirement that it cites the paper "An iterative image registration technique with an application to stereo vison."[1] written by BD. Lucas & T. Kanade and give some insight of something that has happened in the field of optical flow since their paper was published in 1981. Looking through scientific papers citing this paper I found this article, "High Accuracy Optical Flow Estimation Based on a Theory for Warping"[2] discussing a method for performing higher accuracy optical flow estimation. In this paper I will attempt to explain what this article is about, how the proposed technique works, and how it affects the result, e.g. improving the optical flow estimation.

2 The Model

The solution here uses a slightly different model from the one we know of, that applies multiple constraints on our view:

- Grey Value Constancy Assumption, using the original non-linear version due to the linear model falling short when the image changes non-linearly.
- Gradient Constancy Assumption, which assumes the gradient of a grey value does not change as a result of displacement.
- Smoothness Assumption, which attempts to mitigate issues that can arise when a given gradient appears to "vanish" between frames (Think of what happens to small details in an image when you zoom out, or a panning shot with a chain link fence).
- Multiscale Approach, when attempting to track displacements larger than one pixel per frame it can be useful to apply this technique whereby you downsample the image to a lower resolution, solve the problem in that lower scale, and then use that "coarse" solution to refine our answer to the original problem.

Using these assumptions, they derive an energy functional where $x := (x, y, t)^T$, $w := (u, v, 1)^T$ and Ψ is a concave function to prevent outliers from influencing the result too much:

$$E_{Data}(u, v) = \int_\Omega \Psi(|I(x + w) - I(x)|^2 + \gamma|\nabla I(x + w) - \nabla I(x)|^2)dx \tag{1}$$

And a smoothness term to penalize the variations in the flow field:

$$E_{Smooth}(u, v) = \int_\Omega \Psi(|\nabla_3 u|^2 + |\nabla_3 v|^2)dx \tag{2}$$

They then weigh these together:

$$E(u, v) = E_{Data} + \alpha E_{Smooth} \tag{3}$$

The next step is then to find the functions u and v to minimise $E(u, v)$.

3 Numerical Approximation

$E(u, v)$ is a highly nonlinear function, which makes minimisation non-trivial. Hence they decide here to convert the problem into a linear one. The first step is to approach the solution via numerical approximation, using a multiscale approach with an arbitrary factor $\eta \in (0, 1)$, using a complete pyramid of images starting with the lowest resolution and working upwards to more accurate, finer approximates. I'm not entirely sure how the following steps work, but the point is that they start with a highly non-linear function, and through various steps arrive at a linear system of equations:

$$\begin{aligned} 0 = (\Psi')^{k,i}_{Data} * (I^k_x(I^k_z + I^k_x du^{k,l+1}) \\ +\gamma I^k_{xx}(I^k_{xz} + I^k_{xx} du^{k,l+1} + I^k_{xy} dv^{k,l+1}) + \gamma I^k_{xy}(I^k_{yz} + I^k_{xy} du^{k,l+1} + I^k_{yy} dv^{k,l+1})) \\ -\alpha div((\Psi')^{k,l}_{Smooth} \nabla_3(u^k + du^{k,l+1})) \end{aligned} \tag{4}$$

that can be solved using common numerical methods such as Gauss-Seidel or SOR iterations.

4 Relation to Warping

So how does warping relate all of this? The authors of this paper establish this relation by first assuming a simplified model, where by setting $\gamma = 0$ in 1 restrict us to the grey value constancy model and assume solely spatial smoothness.
With these restrictions in place, 4 can be rewritten in a much simpler way.
Only the increments du and dv between the first and second images are estimated, and the same increments appear in the outer fixed point iterations to resolve the nonlinearity of the grey value constancy assumption. This, according the authors, *shows that the warping technique implements the minimisation of a non-linearised constancy assumption by means of fixed point iterations on w.*

5 Evaluation

Here the authors discuss the performance of their solution, by comparing it to the results given by contemporary literature techniques for solving the same problem, using both synthetic and real-world image data. For the synthetic data, this method was shown be over twice as effective compared to prior methods at reducing the angular error of the motion estimates, and for the first time recording an average angular error (AAE) of $< 1°$ for the famous *Yosemite* sequence.
Not only does the technique preserve the discontinuity of the two types of motion in the image, it also estimates the movement of the clouds accurately. This, according to the authors, because of the assumptions stated in the energy functional1: The smoothness assumption allows discontinuities, and the gradient constancy assumption handles brightness changes, like what might typically happen with image sequences involving moving clouds.
To test the robustness of the method, they did two additional tests:
The first where gaussian noise of mean zero and of different standard deviaions were introduced to the sequence. The result of this test showed one thing very clearly: This method performs really well, even with a lot of noise present, where even with a standard deviation of 40 this method still managed to outperform all other techniques operating on an image sequence *without* noise.
The second, where they tweak the parameters of their method (The weight γ, the smoothness parameter α, and the standard deviation σ (Assuming an image has been preprocessed by a gaussian convolution with standard deviation σ)), they computed results with parameter deviations as large as a factor 2 in either direction from the optimal settings. Even with very large parameter changes, the method still produced excellent results.
Not only is the solution accurate, it is also fast, due to the implicit minimisation scheme used, capable of arriving at a good-enough answer with very few outer and inner fixed point iterations, and on a 3GHz Pentium 4 processor (state of the art consumer hardware of the time), capable of completing the calculations in seconds.
Applying the solution to a real-world image sequence, the method was shown capable of providing very realistic results, robust against the interlacing artifacts present in the sequence, and the flow boundaries makes the results directly useful for segmenting an image of vehicles to directly tell diferent vehicles apart.

6 Conclusion & Reflection

In this paper, the authors show a few important points: While brightness and gradient constancy assumptions and spatio-temporal smoothness constraints are all concepts that have proved useful before, this paper shows that the combination of these assumptions into a single method outperforms all previous methods up to the point when this paper was originally published.
The reason for this performance is cited as being due to the energy functional containing *non-linearised* data, and postponing all linearisations to the numerical scheme. Linearisation in the model, according to

the authors, immediately compromises the overall performance of the system, while linearisation in the numerical scheme helps to improve convergence towards the global minimum, i.e. reach the answer faster.
This strategy of transparent continuous modeling in conjunction with consistent numerical approximations, they hope, show that excellent performance and deeper theoretical understanding are not contradictive, just two sides of the same coin.

While I can't claim a perfect understanding of the paper just yet, there is atleast one thing about the method demonstrated that is important to note:
The use of a multiscale approach to approximate a coarse answer, and then through iteration, reach a more accurate answer, looks like a very useful technique that isn't limited to the subject of optical flow estimation; It should be useful in any situation where a problem needs to be minimised.
Given the age of this paper (It's over 10 years old already!), it would not at all be surprising if the method described had already been deprecated by techniques both faster and more accurate. But even so, this was an important milestone on the subject of optical flow estimation that, together with advances in micro controller technology, has enabled higher accuracy motion detection and estimation in realtime without requiring a super computer to perform the calculations.

References

[1] Bruce D. Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vison., 1981.

[2] Thomas Brox, Andrés Bruhn, Nils Papenberg, and Joachim Weickert. High accuracy optical flow estimation based on a theory for warping*, 2004.