Unifying Low-Level Vision

Erik G. Learned-Miller
Department of Computer Science
University of Massachusetts, Amherst
Amherst, MA 01003

May 12, 2011

This white paper supports the goal of **establishing Computer Vision as a coherent intellectual discipline**[1] by suggesting a specific agenda for the unification of many low-level vision principles, algorithms, and data structures. Our goal is to identify a set of highly related low-level vision problems, define their common structure, and establish a coherent intellectual discipline around the shared structures. We believe this can unify and simplify the understanding, teaching, and quality of low-level computer vision as a science.

There is a set of low-level vision tasks that are closely related and share a small set of common principles and yet, are developed mostly separately in the literature. These tasks are background subtraction, tracking (both near-field and far-field), stereo vision, image stitching, image registration (both medical and non-medical), object recognition of "low variance" object classes (such as rigid objects), optical flow, and general alignment algorithms. The literature on background subtraction, for example, has fewer references to tracking than one would expect given their closely related nature.

At the heart of these tasks is a *low-level image representation* and an *image comparison function* giving some notion of similarity or distance between two images or image patches. There has been tremendous innovation in developing low-level image representations and image comparison functions. Some examples include SIFT descriptors [6], HoG descriptors [3], geometric blur [1], the pyramid match kernel [5], mixtures of Gaussians at each pixel [8], non-parametric distributions at each pixel [4], affine invariant descriptors [7], and many others.

We would like to revisit these descriptors and ask the question, "What properties do we want such a descriptor, and its associated comparison function, to have?" Developing primitive functions by specifying their requirements first has a history of success in the mathematical sciences. Two salient examples include Shannon's definition of information entropy and Einstein's development of Special and then General Relativity. Shannon required that a measure of information be continuous, symmetric, additive, and should be highest when all outcomes are equally likely. These requirements constrain the definition of entropy up to a positive constant (the units). Similarly, Relativity essentially follows from the simple *Principle of Equivalence* of reference frames.

In computer vision, this approach is highlighted by such foundational works as Canny's edge detector [2], defining the optimal edge detector with respect to certain requirements. In a similar spirit, we would like to return to the requirements of low-level descriptors and comparison functions in view of developments from the last 20 years.

What should the requirements of low-level representations and comparison functions be? Some candidates include

- robustness to image noise,
- robustness to small misalignments,
- a multi-scale or multi-resolution aspect,
- absence of "hard bins", which introduce oversensitivity to position,
- "denseness", the property that every location in the image contributes information to the descriptor,

[1] This is key objective number 2, from the Frontiers in Computer Vision white paper by Yuille and Oliva.

- smooth degradation as a function of spatial location,
- smooth degradation as a function of scale,
- smooth degradation as a function of brightness,
- ability to rapidly adapt parameters to the image at hand,
- compatibility with different low-level feature types, such as color and edges,
- having a clear (possibly probabilistic) interpretation,
- having a simple way to represent uncertainty in measurement or value of primitive features.

By examining desirable properties of representations and comparison functions, an interesting pattern emerges: all of the desirable properties appear in some descriptors, but no descriptor has all of these properties. For example, geometric blur exhibits smooth degradation as a function of spatial location but is not dense and doesn't adapt to the image at hand. SIFT is robust to small misalignments but has hard bins and fixed parameters. The pyramid match kernel is multi-scale, but has discrete bins and is not dense.

We promote the idea that all of the descriptors and comparison methods mentioned can be understood as **distribution fields**, i.e. probability distributions over primitive features at each point in an image. Stauffer and Grimson's backgrounding model is a distribution field: each pixel distribution is modeled as a mixture of Gaussians. SIFT is an approximate distribution field: at each of 16 different image patch locations, a pseudo-probability distribution[2] over edge energies is defined. By describing descriptors in this way, two things are achieved. First, the differences between them are clarified. Second, it suggests simple methods for achieving new descriptors that share the properties of multiple previously successful descriptors. We now describe a subtask shared by many low level vision problems that, in our opinion, can act as a generic way to assess the descriptor outside the context of the thorny details of each individual problem.

A common subtask: Basin of attraction studies

One of the nice pieces of recent scientific work in the computer vision community has been a careful and thorough analysis of the problem of invariant descriptors [7]. Since many computer vision algorithms can benefit from invariant descriptors, this work has been invaluable in improving, unifying, and simplifying computer vision. We would like to propose another canonical problem that we believe is shared by the low-level vision problems identified above. We call it the **basin of attraction problem,** defined as follows.

Given an image I and a small image patch J, either from the same image or a different image, find the patch of I (under some set of transformations) that best matches J through **local search**. That is, we wish to define an experiment in which the patch J is "near" to its ideal location in image I, and follow the gradient of a comparison function to its local minimum. We can then ask the question, "From how far away, on average, does a patch find its globally optimum match." The set of locations from which a patch can follow a gradient to its optimum position is the **basin of attraction** of the optimum. Clearly it is a function of the low-level representation and comparison function. We believe this is an important focus for the science of computer vision because it isolates a key element of many of the low-level algorithms while separating it from problem-specific details that differ among the problems.

Our central thesis is this. **Representations and comparison functions that lead to large basins of attraction are highly desirable for backgrounding, tracking, optical flow, and many other low level vision problems.** Even when non-local search, such as keypoint based matching, is used, matches can be refined after the keypoint location is found. We call such matches **sharpening matches**, and they have the potential to improve performance on many low-level problems.

We propose the following specific agenda items for part of the workshop:

- What are the core shared concepts in the low-level tasks identified in this paper?
- Discuss the importance of local matching to computer vision. When is global matching (keypoint methods or exhaustive search) better? When is it inferior?[3]

[2]The only difference between the vectors of edge strengths in SIFT is that they are normalized using the L2 norm instead of the L1 norm, which would result in valid probability distributions.

[3]Szeliski has an interesting discussion of these issues in his recent book [10, 9].

- Clearly introduce the basin of attraction problem and present results from our studies of many common descriptors. Is this an important shared problem in Computer Vision?
- Discuss desirable properties of descriptors and comparison functions.
- Discuss each of the most popular descriptors and comparison functions and how they can be related through their representation as distribution fields.
- Define the notion of a sharpening match, and discuss its relevance.

References

[1] Alexander C. Berg and Jitendra Malik. Geometric blur for template matching. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2001.

[2] J Canny. A computational approach to edge detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 8:679–698, November 1986.

[3] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 886–893, June 2005.

[4] Ahmed Elgammal, Ramani Duraiswami, David Harwood, Larry S. Davis, R. Duraiswami, and D. Harwood. Background and foreground modeling using nonparametric kernel density for visual surveillance. In *Proceedings of the IEEE*, pages 1151–1163, 2002.

[5] Kristen Grauman, Trevor Darrell, and Pietro Perona. The pyramid match kernel: Efficient learning with sets of features. *Journal of Machine Learning Research*, 8:2007, 2007.

[6] David Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 2003.

[7] K. Mikolajczyk and C. Schmid. Comparison of affine-invariant local detectors and descriptors. In *European Signal Processing Conference*, 2004.

[8] Chris Stauffer and W. Eric. L. Grimson. Learning patterns of activity using real-time tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):747–757, 2000.

[9] Richard Szeliski. Image alignment and stitching: A tutorial. Technical Report MSR-TR-2004-92, Microsoft Research, 2006.

[10] Richard Szeliski. *Computer Vision: Algorithms and Applications*. Springer, 2011.