Lecture 5: Image Statistics and Weak Membrane Models

- Statistics of the derivatives of images. *naturalness*
- Weak Membrane Models. Mumford and Shah. Rudin, Osher, Fatemi.

Statistics of Image Derivatives

- The statistics of image derivatives are extremely consistent for natural images. Some researchers call this a *naturalness prior*. These statistics differ greatly from those of random noise images (e.g., a flicking TV set).
- To explore this, differentiate an image $I(x, y)$ – to obtain dI/dx – and compute its histogram. This has a Laplacian distribution $p(x) = \frac{1}{Z(k)} \exp\{-k|x|\}$, where k is a positive constant and $Z(k)$ normalizes the distribution.
- Note that if the derivatives were normally distributed, then the plots would look like a Gaussian and the "tails" would fall off rapidly, like $\exp\{-(1/2)x^2\}$. But instead they fall off much more slowly – the Gaussian distribution is not *robust* enough to deal with this this data (it under-estimates the changes of rare events – e.g., the black swans that arguably caused the great recession).
- Intuitively, dI/dx is large at edges in the image (at boundaries of objects or at sharp texture boundaries) and tends to be smaller elsewhere (note: but *texture* contains many small edges.). So this empirical finding (discovered in the 1990's) suggests that the image derivatives are small at many positions in the image, in other words images are *piecewise smooth*. This related to the *weak membrane* models developed in the 1980's (see later this lecture). Membranes are smooth (e.g., soap bubbles) so weak membrane are piecewise smooth (see also markov random field models later in the course).

Statistics of Generalized Image Derivatives: Green 1

- Researchers found that the statistics of higher order derivatives of images also followed Laplacian distributions. This is less easy to reconcile with the weak membrane models. Higher order derivatives correspond to longer range spatial interactions.
- Intuitively, first order derivative can be approximated by the difference between neighboring points on an image lattice, the second order derivative requires considering the intensity values at three (or more) neighboring points, and in general the n^{th} order derivative requires considering the intensity value at n neighboring points.
- Researchers (e.g., M. Green) did studies on generalized derivatives: $X_A = \sum_{i=1}^{n} a_i X_i$ where $\sum_{i=1}^{n} a_i = 0$. He found that these also obey Laplacian distributions. This is inconsistent with weak membrane models.
- M. Green (who is a pure mathematician) took this idea to its logical extreme. If generalized derivatives obeyed this property – i.e. had Laplacian distributions $f(X) = \frac{\beta}{2} \exp\{-\beta|x|\}$ for some unknown β. — $\forall \{a_i, ..., a_n\} \ \text{s.t.} \ \sum_{i=1}^{n} a_i = 0$. He calls this property *differentially Laplacian*. This assumes that the Laplacian parameter β is the same for all choices of $\{a_1, ..., a_n\}$, and Green provides evidence for this.

Statistics of Generalized Image Derivatives: Green 2

- Green shows that this property can be exploited to define probability distribution $P(X_1, ..., N_n)$ on the pixel intensities within an 8×8 patch. He first observes that the new set of variables $Y_1, ..., Y_{n-1}$, defined by $Y_i = X_i - X_n$, will obey the stronger *Linear Laplacian constraint* $Y_A = \sum_{i=1}^{n-1} a_i Y_i \quad \forall \{a_i, ..., a_{n-1}\}$ will obey the same Laplacian distribution (i.e. without requiring the constraint $\sum_{i=1}^{n-1} a_i = 0$).
- He shows that the probability distribution for any variables which are linearly Laplacian must obey: $\int p(Y_1, ..., Y_{n-1}) \exp\{-i \sum_{i=1}^{n-1} Y_i \omega_i\} d\vec{Y} = \frac{2}{||\omega||_B^2 + 2}$, where $||\omega||_B^2 = \sum_{i,j=1}^{n-1} B_{ij}\omega_i\omega_j$ where B_{ij} is the correlation function between Y_i and Y_j, i.e. $B_{ij} = \int Y_i Y_j P(Y_1, ..., Y_{n-1}) d\vec{Y}$.
- This specifies a distribution for $X_1, ..., X_N$ by $P(X_1, ..., X_{n-1}|X_n)P(X_n)$. with $P(X_1, ..., X_{n-1}|X_n) = P(X_1 - X_n, ..., X_{n-1} - X_n)$. To go further, we must know the correlation functions B_{ij} of Y_i and Y_j. These can be expressed in terms of the correlation functions between $X_i - X_n$ and $X_j - X_n$, in other words $< X_i X_j > - < X_i, X_n > - < X_j - X_n > + < X - n, X_n >$.
- The correlation functions between intensity values X_i, X_j have been measured and generally obey a fall-off rule: $< X_i X_j >= (1 + \gamma d_{ij})^{-\alpha}$, where α, γ are constants, and d_{ij} is the distance between the pixels i and j.

Statistics of Generalized Image Derivatives: Green 3

- There are several points to make.
- Firstly, these statistics show that there are many image regularities which are not captured in the statistics of the first order derivatives of the images. This implies there is longer range structure (we will return to this when we discussion variational models and markov random fields).
- Secondly, these differentiable Laplacian statistics are independent of linear transformations on the images. These linear transforms include scaling the image by multiplying it by a constant and, more importantly, rescaling by averaging the image within boxes This show that these statistics are independent of scale.
- Thirdly, Green argues that similar statistics occur for many physical stimuli and not just images (relates to the scaling properties).
- But what does this mean for computer vision? *Naturalness Priors must be based on high order derivative statistics of images* (can be used to regularize images – helpful for some deep network applications). The weak membrane models (following) are problematic because they ignore these higher order derivatives. Image patch models, however, are perhaps better because they capture non-local correlation.

Weak Membrane: Mumford and Shah (1)

- Mumford and Shah formulated image segmentation of a domain D as the minimization of a functional $E[J, B]$. The input I is an image, The output $(\hat{J}, \hat{B}) = \arg\min E[J, B]$ is a smoothed image $\hat{J}$ and the position $\hat{B}$ of the boundaries that separates D into subdomains $D = \bigcup D_i$, with $D_i \bigcap D_j = 0 \ i \neq j$ (and $B = \bigcup \partial D_i$) (i.e. b specifies the positions of a one-dimensional set of points). $E[J, B]$ is called a functional because the argument is a function $J(\vec{x})$.
- $E[J, B] = C \int d\vec{x}(I(\vec{x}) - J(\vec{x}))^2 + A \int_{D/B} \vec{\nabla} J(\vec{x}) \cdot \vec{\nabla} J(\vec{x}) d\vec{x} + B \int_B ds.$
- The first term ensures that the the smoothed image J is similar to the input I. The second term ensures that J has small gradient $|\vec{\nabla} J|$ (i.e. J is smooth) except across boundaries B. The third term penalizes the length of the boundaries ($\int_B ds$ is a one-dimensional integral). Intuitively, it tries to smooth the image I except at places where the image gradient $|\vec{\nabla} I|$ is too high – where it cost less energy to insert a boundary/edge.
- This model exploits both edge and regional cues: (i) *edge cues:* it tries to insert boundaries at places where the gradient of the image $I(\vec{x})$ is large, and (ii) *regional cues:* it tries to group pixels which have similar intensities into regions. It is a type of week-smoothness or weak-membrane model.

Weak Membrane: Mumford and Shah (2)

- The Mumford and Shah model is of considerable historical and mathematical interest. Mathematically, it was non-trivial to prove that the energy functional had well defined minima (Ambrosio and Torterelli).
- Historically, it was one of the three classic weak-smoothness/weak-membrane models proposed for image segmentation. The two other (Geman and Geman – Blake and Zisserman) were formulated in terms of Markov Random Fields). All these models (invented independently in the early 1980's) combined edge and regional cues.
- Mumford and Shah has practical limitations. The energy functional is non-convex so it is not easy to specify algorithms that will minimize it (and impossible to specify an algorithm that converges to the global optimum.
- Like other weak-membrane models, Mumford and Shah relies on first order derivatives ((or first order differences, when converted to a discrete lattice, hence nearest neighbor interactions for Markov Random Fields). This means that it cannot capture non-local statistics and hence is not a very accurate model of natural images.

Weak Membrane: Mumford and Shah (3)

- Ambrosio and Torterelli. Define a functional $E[J,z;\epsilon] = C\int (J(\vec{x}) - I(\vec{x}))^2 d\vec{x} + A\int z(\vec{x})|\vec{\nabla}J(\vec{x})|^2 d\vec{x} + B\int \{\epsilon|\vec{\nabla}z(\vec{x})|^2 + \epsilon^{-1}\phi^2(z(\vec{x}))\} d\vec{x}$.
- where $\epsilon > 0$ is small parameter and $\phi(z)$ is a potential function. A choice for $\phi(z) = (1-z)/2$ for $z \in [0,1]$. The edge set B will be the set of points z such that $\phi(z) \approx 0$ (i.e. $z \approx 1$).
- It can be shown that in the limit as $\epsilon \mapsto 0$ that this energy functional is equivalent to Mumford and Shah.
- Steepest descent can be performed on $E[J,z;\epsilon]$ with respect to J and z while gradually decreasing ϵ. This will converge to a minimum of Mumford and Shah.

Weak Membrane: Rudin-Osher-Fatemi (1)

- This models is not strictly a weak membrane model, but it has many of their good properties and some of their bad properties (like having local interactions). It also has the big advantage that it is convex. Hence applied mathematicians can develop efficient algorithms for finding its global minimum. For these reasons it was very effective for image denoising (until replaced by patch-based methods – e.g., dictionaries – which were able to capture longer-range interactions.
- The Rudin-Osher-Fatemi model takes the form: $E[J; I] = \int_D |\vec{\nabla} J| d\vec{x} + \frac{\lambda}{2} \int_D (J(\vec{x}) - I(\vec{x}))^2 d\vec{x}$. Minimizing this with respect to J gives a smoothed image. Thresholding the derivatives of J gives the edges.
- This functional is convex since it consists of an $L1$ norm tern plus a quadratic term. Both terms are convex, so their sum is also convex.
- Unlike the weak membrane models, this model does not decompose the image into a sum of disjoint regions. It does smooth images across boundaries (if the boundaries are defined by thresholding the gradients of J). This is slightly ugly. But is is a necessary price for ensuring that the energy functional is convex.

Weak Membrane: Rudin-Osher-Fatemi (2)

- The Rudin-Osher-Fatemi model can be reformulated to have variable that are like edges.
- $E[J, z] = \frac{1}{2} \int_D \{z|\vec{\nabla} J|^2 + z^{-1}\} d\vec{x} + \frac{\lambda}{2} \int_D (J - I)^2 d\vec{x}$.
- We can solve for $z = 1/|\vec{\nabla} J|$ (differentiate $E[J, z]$ with respect to z) and substitute back to obtain the Rudin-Osher-Fatemi model. We can interpret z as a measure of edgeness – $z = 0$ indicates an edge.
- Alternative minimization can be used to minimize $E[J, z]$. Minimizing $E[J, z]$ with respect to z yields $z^{t+1} = 1/|\vec{\nabla} J^t|$. Minimizing $E[J, z]$ with respect to z requires solving: $-2\vec{\nabla}\{z^t \vec{\nabla} J^{t+1}\} + \lambda(J^{t+1} - I) = 0$, which has a unique solution (since $E[J, z]$ is a convex function of J if z is fixed).
- This alternative minimization reduces to the well-known *lagged-diffusion* model: $-\vec{\nabla} \cdot \{|\vec{\nabla} J^t|^{-1} \vec{\nabla} J^{t+1}\} + \lambda(J^{t+1} - I) = 0$.
- Note: should have a link to level-sets, split-Bregman, and other algorithms for this type of problem.

Convexity and Steepest Descent

- An energy functional (or function) $E[J; I]$ is convex if for all $0 \leq \alpha \leq 1$ and any J_1, J_2 we have $\alpha E[J_1, I] + (1-\alpha) E[J_2, I] \geq E[\alpha_1 J_1 + (1-\alpha_1) J_2]$. This is equivalent to the condition that the Hessian of $E[J; I]$ – the second order derivatives $\frac{\delta^2 E}{\delta J^2}$ – is positive semi-definite.
- Steepest descent updates J in the direction of the gradient $-\frac{\partial E}{\partial J}$ (i.e. downhill in $E[J; I]$. This is guaranteed to reduce the energy: $\frac{dJ}{dt} = -\frac{\partial E}{\partial J}, \mathrm{Implying} \frac{dE}{dt} = -\frac{\partial E}{\partial J}\frac{dJ}{dt} = -\frac{\partial E}{\partial J}\frac{\partial E}{\partial J} \leq 0$.
- If $E[J]$ is convex and bounded below (e.g., $E[J] \geq 0$) then $E[J]$ has a unique minimum (which is global) and steepest descent is guaranteed to find it. There are non-convex functions which only have a singe (global) minimum. But convexity is the only criterion that can be easily checked to guarantee that an energy function has a unique minimum.
- Concave functions are the opposite of convex functions (i.e. if $f(x)$ is convex then $-f(x)$ is concave, and vice versa).Surprisingly most functions can be decomposed into the sum of convex and concave terms.

Variational Bounding and CCCP (1)

- Steepest descent methods need to be discretized in time to be implemented by computers. We must convert the update equation $\frac{d\vec{x}}{dt} = -\frac{\partial f}{\partial \vec{x}} = -\vec{\nabla} f(\vec{x})$ into a discrete update rule: $\vec{x}_{t+1} = \vec{x}_t - \Delta \vec{\nabla} f(\vec{x}(t))$. But finding a good value for Δ is difficult. Too small a values makes steepest descent go very slowly. But too big a value may prevent the algorithm from converging.
- Discrete iterative algorithms are an alternative. Variational bounding proceeds by obtaining a sequence of bounding functions $E_B(\vec{x}, \vec{x}_n)$ where $\vec{x}_n$ is the current state. The bounding functions must obey: $E_B(\vec{x}, \vec{x}_n) \geq E(\vec{x}), \forall \vec{x}, \vec{x}_n$ and $E_B(\vec{x}_n, \vec{x}_n) = E(\vec{x}_n)$.
- Then the algorithm $\vec{x}_{n+1} = \arg\min_{\vec{x}} E_B(\vec{x}_n)$ is guaranteed to converge to a minimum of $E(\vec{x})$. The algorithm can make large moves from $\vec{x}_n$ to $\vec{x}_{n+1}$.

Variational Bounding and CCCP (2)

- A special case of this approach is called CCCP. Decompose the function $E(\vec{x})$ into a concave $E_c(\vec{x})$ and a convex part $E_v(\vec{x})$ so that: $E(\vec{x}) = E_c(\vec{x}) + E_v(\vec{x})$.
- Then the update rule $\vec{\nabla} E_v(\vec{x}_{n+1}) = -\vec{\nabla} E_c(\vec{x}_n)$ is guaranteed to decrease the energy. This can be shown directly, or follows from variational bounding where $E_B(\vec{x}, \vec{x}_n) = E_v(\vec{x}) + E_c(\vec{x}_n) + (\vec{x} - \vec{x}_n) \cdot \vec{\nabla} E_c(\vec{x}_n)$.
- It can be shown that many existing discrete iterative optimization algorithms can be re-expressed as CCCP or variational bounding (sometimes by performing changes of variables).
- Even Steepest Descent can be derived as a special case. Express $E(\vec{x}) = E(\vec{x}) + \lambda/2|\vec{x}|^2 - \lambda/2|\vec{x}|^2$. If we make λ sufficiently large, then $E(\vec{x}) + \lambda/2|\vec{x}|^2$ will be convex and $-\lambda/2|\vec{x}|^2$ will be concave. Applying CCCP we rederive iterative steepest descent (with Δ depending on λ).

To Do

- Add calculus of variations.
- Maybe add Legendre transform.