Lecture 4: Image Statistics and Weak Membrane Models

- Statistics of the first order derivatives of images. *naturalness*
- Weak Membrane Models. Mumford and Shah. Geman and Geman, Blake and Zisserman. Rudin, Osher, and Fatemi.
- Statistics of higher order derivatives.

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.
- Intuitively, dI/dx is large at at a small faction of image pixels, the *edges in the image* (e.g., at boundaries of objects) and tends to be smaller elsewhere (except in textures which contain many small edges).

Statistics of Image Derivatives

- If the derivatives were normally distributed by a Gaussian, then the "tails" would fall off rapidly, like $\exp\{-(1/2)x^2\}$. But instead they fall off.
- This rapid fall off s to a bad property of Gaussians. They are not *robust* to outliers. Estimates of the mean and variance of data $\{x_i\}$ can be strongly influenced by a only few *outliers*. These are data points x whose values differ greatly from the majority of the $\{x_i\}$.
- There is a research field called Robust Statistics (PJ Huber). This is important to know and is the basis of some recent work on the domain transfer peoblem of neural networks.

Energy Function Models and Image Statistics (1)

- These statistical finding seem, at first sight, to justify a class of *weak membrane*, or *weak smoothness* models. The term *membrane* is used because physical Membranes (e.g., soap bubbles) has smooth surfaces.
- There are also probabilistic weak smoothness models (Geman and Geman) and related Blake and Zisserman. The maximum entropy principle can learn probability models from image statistics. And particularly derive weak smoothness models from image deriavtive statistics (See later)

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 $D_i \bigcap D_j = 0$ $i \neq j$.
- $\hat{B} = \bigcup \partial D_i$, i.e. $\hat{B}$ specifies the boundaries between the domains.

Weak Membrane: Mumford and Shah (1)

- $E[J, B]$ is called a functional because the argument is a function $J(\vec{x})$. By contrast, the argument of a function $J(\vec{x})$ is a variable $\vec{x}$.
- To "differentiate" a functional you use the *calculus of variations* instead of standard differentiation (beyond scope).
- $E[J, B] =$
  $C \int d\vec{x}(I(\vec{x}) - J(\vec{x}))^2 + A \int_{D/B} \vec{\nabla} J(\vec{x}) \cdot \vec{\nabla} J(\vec{x}) d\vec{x} + B \int_B ds.$

Weak Membrane: Mumford and Shah (1)

- 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.

Weak Membrane: Mumford and Shah (1)

- This model exploits both edge and regional cues.
- *Edge cues:* it tries to insert boundaries at places where the gradient of the image $I(\vec{x})$ is large, and
- *Regional cues:* It tries to group pixels which have similar intensities into regions. This is why it is called a 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.

.

Weak Membrane: Mumford and Shah (3)

- 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.
- We present an algorithm by Ambrosio and Torterelli.

Weak Membrane: Mumford and Shah (3)

- Define a functional $E[J, z; \epsilon] = C \int (J(\vec{x}) - I(\vec{x}))^2 d\vec{x} + A \int z(\vec{x}) |\vec{\nabla} J(\vec{x})|^2 d\vec{x} + B \int \{\epsilon |\vec{\nabla} z(\vec{x})|^2 + \epsilon^{-1} \phi^2(z(\vec{x}))\} d\vec{x}$.
- $\epsilon > 0$ 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. $z \approx 1$).
- Steepest descent can be performed on $E[J, z; \epsilon]$ with respect to J and z while gradually decreasing ϵ converging to a minimum of Mumford and Shah.

Weak Membrane: Rudin-Osher-Fatemi (1)

- This models has many good properties of Weak Membrane models but the big advantage of minimizing a convex energy functional.
- 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.

Weak Membrane: Rudin-Osher-Fatemi (1)

- 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. But boundaries can be found by thresholding.

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.

Weak Membrane: Rudin-Osher-Fatemi (2)

- 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: $-\vec{\nabla} \cdot \{|\vec{\nabla} J^t|^{-1} \vec{\nabla} J^{t+1}\} + \lambda(J^{t+1} - I) = 0$.

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 $\alpha E[J_1, I] + (1 - \alpha)E[J_2, I] \geq E[\alpha_1 J_1 + (1 - \alpha_1)J_2]$. 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: $\frac{dJ}{dt} = -\frac{\partial E}{\partial J}, \text{Implying} \frac{dE}{dt} = -\frac{\partial E}{\partial J}\frac{dJ}{dt} = -\frac{\partial E}{\partial J}\frac{\partial E}{\partial J} \leq 0.$

Convexity and Steepest Descent

- 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 can be found by steepest descent. There exist non-convex functions (e.g., a golf hole) which only have a singe minimum. But convexity is the only condition which can be easily checked.
- 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.

Convexity and Steepest Descent

- Steepest descent must be discretized for implementation. Convert 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 suitable Δ is difficult. Too small makes steepest descent slow. Too big makes the algorithm unstable.
- There is a large literature on variants of steepest descent. Probably the best known is Newton-Raphson which uses second order derivatives as well as first order. These are beyond the scope of the course.

Variational Bounding and CCCP (1)

- 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 $E(\vec{x})$. It can make a big 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 $E_B(\vec{x}, \vec{x}_n) = E_v(\vec{x}) + E_c(\vec{x}_n) + (\vec{x} - \vec{x}_n) \cdot \vec{\nabla} E_c(\vec{x}_n)$.

Variational Bounding and CCCP (2)

- It can be shown that many existing discrete iterative optimization algorithms can be re-expressed as CCCP or variational bounding (may need 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).

Stats of Generalized Image Derivatives

- The weak membrane models rely on first order derivatives. These are first order differences when converted to a discrete lattice and local spatial interactions. The models can be justified by the experimental findings of the statistics of the first order derivatives of natural images.
- But researchers studied the statistics of higher order derivatives of images, i.e. longer range interactions. These also followed Laplacian distributions and hence inconsistent with weak membrane.

Stats of Generalized Image Derivatives

- Intuitively, first order derivative can be approximated by the difference between neighboring points on an image lattice and 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.

Stats of Generalized Image Derivatives

- M. Green provided evidence that all generalized derivatives were *differentially Laplacian* $f(X) = \frac{\beta}{2} \exp\{-\beta|x|\}$ for some unknown β. — $\forall\{a_i, ..., a_n\}$ s.t. $\sum_{i=1}^{n} a_i = 0$ with the same parameter β is the same for all choices of $\{a_1, ..., a_n\}$..
- He used this conjecture to show the 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 $\sum_{i=1}^{n-1} a_i = 0$).

Stats of Generalized Image Derivatives

- Green 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 $P(X_1, ..., X_{n-1}|X_n) = P(X_1 - X_n, ..., X_{n-1} - X_n)$.

Stats of Generalized Image Derivatives

- To go further, we must know the correlation functions B_{ij} of Y_i and Y_j. These are given by the correlation functions between $X_i - X_n$ and $X_j - X_n$;
  $< X_i X_j > - < X_i, X_n > - < X_j, X_n > + < X_n, X_n >$.
- 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.

Stats of Generalized Image Derivatives

- Greens studies (and others by Mumford) make important points.
- (1) There are image regularities which are not captured in the statistics of the first order derivatives of the images. This implies there is longer range structure not captured by weak membrane models. Suggesting alternatives, like mixture of Gaussian patches may be better.

Stats of Generalized Image Derivatives

- (2) These differentiable Laplacian statistics are independent of linear transformations on the images. These include scaling the image by multiplying it by a constant and by rescaling it by averaging the image within boxes *This show that these statistics are independent of scale*.
- (3) Green argues that similar statistics occur for many physical stimuli and not just images (relates to the scaling properties).