Analysis by Synthesis: Overview

- Helmholtz and Gregory argued that vision was an inference process that combined direct stimulus input (images) with top down processes involving knowledge and memory.
- One way to formalize this in in terms of Bayesian Inference. Suppose I and W specify the image and the state of the world (respectively) then we seek to estimate W from the posterior $P(W|I)$. This posterior is the product of the likelihood term $P(I|W)$ with the prior $P(W)$ (normalized by $P(I)$). The prior $P(W)$ specifies the prior knowledge about the world, while $P(I|W)$ gives the consistency of the image with the world state (see Pawan Sinha's figure which analyzes how humans perceive the Necker cube from a Bayesian perspective).
- Bayes motivates the concept of *analyze by synthesis*. To *synthesize* an image, we first sample W from $P(W)$, and then sample I from $P(I|W)$ (this sampling can be done by MCMC in principle). To be able to synthesize realistic images requires specifying realistic models $P(I|W)$ and $P(W)$ (which could be done by computer graphics).
- To *analyze* an image I we need to invert the synthesis process and find the world state W which is most likely to have generated the image I. This is the core idea of *analyze by synthesis*. The main challenges are to obtain realistic models $P(I|W), P(W)$ for generating/synthesizing images and inference algorithms to find $\hat{W} = \arg\max P(W|I)$. (The dumbest inference algorithm is simply to sample $P(I|W)$ from each W until we generate an image that matches the input image).

Analysis by Synthesis: Neuroscience

- In 1991 Mumford conjectured that the human visual system performed analysis by synthesis. He was motivated by two considerations: (i) analysis by synthesis was the right strategy for "solving vision" (based on arguments from Grenander), and (ii) neuroscience studies showed that there are even more feedback connections than feedforward connections (e.g., in the ventral stream). The main idea is that the feedforward path takes the image I as input and activates high level hypotheses W. These high-level hypotheses generate images, by sampling from $P(I|W)$, using the feedback path and these are matched to the input image.
- This contrasts with traditional models of the ventral stream which are typically feedforward (e.g., Fukushima, Marr, Poggio's HMax) and with Deep Networks. Deep Networks learn a discriminative model $P(W|I)$ but have no generative model $P(I|W)$ or prior $P(W)$. Note: it is much easier to learn a discriminative model $P(W|I)$ than a generative model $P(I|W)$ since the set of W is much lower dimensional than the set of images I. Note: there are generative Deep Networks (GANs and auto-encoders) but these have not yet been sufficiently successful for generating realistic image patterns (despite claims, and this could change).
- The role of feedback processing in neuroscience is currently unclear (the ventral stream is so complicated). There are relevant fMRI studies by Kersten and others. There are recent techniques which enable top-down processing to be "switched-off" which offer ways to investigate this in more detail.

Analysis by Synthesis

- The Data-Driven Markov Chain Monte Carlo (DDMCMC) system (Tu and Zhu 2002, Tu, Chen, Yuille, Zhu 2006) is an algorithm for performing analysis by synthesis. It uses bottom-up discriminate methods to active high-level hypotheses which generate images which can be matched to the input (and validated or rejected).
- The specifies the set of images by a grammer whereby an image can be divided into multiple different regions, where each region has a type corresponding to different patterns (e.g., face, text, texture, tec) and parameters of the models. It has generative models for all types of patterns. It also has a (weak) prior on the types of patterns that can occur and their spatial configuration.
- Technically DDMCMC is a Metropolis-Hastings MCMC algorithm. It has a set of proposals $\alpha(.)$ for how to change the estimated state W. These include merging two image regions into a single region, splitting a region into two, assigning/changing the type of a region, changing the parameters of a region, and deforming the boundary shape of a region. This is the *proposal stage* of the algorithm. It then accepts the proposals with a probability which depends on whether making the proposal increases the posterior distribution $P(W|I)$ (with a discount factor depending on the proposal $\alpha(.)$). It can be shown that this algorithm converges to samples from $P(W|I)$

Analysis by Synthesis: Results

- The DDMCMC paper by Tu and Zhu (2002) was state-of-the-art for image segmentation. It had four types of generic pattern models. Proposals included using edge detectors combined with spatial grouping to make proposals for the regions.
- The DDMCMC paper by Tu et al (2006) added models of faces and letters/digits chosen because: (i) reasonable generative models of faces and letters/text existed, (ii) also there were good bottom-up methods for detecting faces/text/digits which could be used to make proposals.
- The Tu et al (2005) paper was conceptual and not evaluated on a benchmarked dataset (no suitable dataset existed). The work showed cooperation (different models combining to explain different parts of the image) and competition (different models competing to explain an image region). It also gave a model of the entire image (every pixel was assigned to a model) which enabled the algorithm to "explain away" subregions which did not fit the model (e.g., the face model didn't include dark glasses, but these could be treated as "generic regions" which could explain why the eyes of the humans were missing).
- Analysis by synthesis remains "ahead of its time". We lack generative models of objects and scene structures (except computer graphics) and the proposals are ad hoc. By contrast, compositional models are roughly generative and gives natural ways to make proposals (by finding elementary components that can be combined together to build larger components).