MRF for:  
- Classification  
- Segmentation

- Fusion: General Remote Sensing & Hyperspectral fusion  
  
My MRF fusion is fusion on: abundances (physical meaning) & classifier output on hyperspectral images  
Davis: on multispectral + he fuses pixel elements with superpixels - objects

Motivation for fusion from:

**Multi-modal obstacle detection paper**

* Due to the difference in their (physical sensing), the detection capabilities of these modalities both complement and overlap each other….
* In contrast, actual sensor fusion provides reduced uncertainty when combining multiple sensors as opposed to applying each sensor individually.
* high-level (late) fusion, integrating information at decision level.
* by probabilistically fusing their equally weighted classffier outputs
* One sensor may recognize class A but confuse B and C, whereas another sensor may recognize C but confuse A and B. By learning this relationship, the sensors can be fused to effectively distinguish all three classes
* works by jointly inferring optimal class labels of 2D segments in images and 3D segments

in corresponding point clouds

* This provides pairwise edges between 2D and 3D segments, thus allowing one modality to correct the initial classification result of the other
* Once initial probability estimates of all 2D and 3D segments have been found and their edges defined, an undirected graphical model similar to the one visualized in Figure 6 can be constructed
* 2D – 3D edges,….terms by introducing costs for assigning different labels.   
  We introduce a cost of assigning corresponding 2D and 3D nodes with different class labels  
  The cost depends on the exponentiated distance between the two neighbors, such that a small distance will incur a high cost and vice versa (Boykov and Jolly 2001; Krahenbuhl and Koltun 2012).
* Beta and gamma represents the importance of the pairwise potentials
* Indicator function (delta function) ensures that the potential is zero for neighboring segments that are assigned the same label.
* However, since the potential concerns different domains (2D and 3D), That is, the cost of assigning x2Di to class A and x3D i to class B might not be the same as the other way around. This allows for interactions that depend on both class label and sensor technology.  
    
  Is my graph cyclic?  
  Since the graph is cyclic, exact inference is intractable and loopy belief propagation is therefore used for approximate inference, while I use GraphCut for approximate inference.
* The same applies at test time for decoding. The decoding procedure seeks to determine the most likely configuration of class labels by minimizing the energy E (x j z). The energy can thus be seen as a cost for choosing the label sequence x given all measurements z.
* The overall results are presented by evaluating the improvement in classification when introducing the fusion algorithm
* Table 2 presents the results for applying the CRF with the three different types of pairwise potentials enabled. Initial, CRF2D, and CRF3D thus refer to single-modality results obtained with the direct output of the initial 2D or 3D classifier and the \smoothed" version of the CRF, respectively. CRF2D-3D additionally introduces sensor fusion by adding edges across the two modalities.
* From Table 2, we see a gradual improvement in classification performance when introducing more terms in the CRF. First, the initial classifiers for 2D and 3D were improved separately by adding spatial links between neighboring segments
* Then, by introducing multi-modal links between 2D and 3D, the performance was further increased.
* the initial classi\_cation of the image was noisy and a\_ected by saturation problems in the raw image. When introducing 2D edges in the CRF (d), most of these mistakes were corrected. Finally, when combined with information from 3D, the CRF was able to correct vegetation and ground pixels around the trailer.
* The **influence on sensor (CRF) fusion** **depends on distance**, since the pixel areas of 3D segments decrease with distance. The error due to calibration increases with distance
* Due to the physics of the camera and the lidar, the two modalities perceive signi\_cantly different characteristics of the environment
* Conclusions:  
  Initial 2D (camera) and 3D (lidar) classi\_ers have been combined probabilistically, exploiting both spatial, temporal, and multi-modal links between corresponding 2D and 3D regions. The method proves that it is possible to reduce uncertainty when probabilistically fusing lidar and camera as opposed to applying each sensor individually.

**Multi-modal Scene Analysis paper**

* This graph encourages each region/node in a modality to leverage the information from its corresponding regions/node in the other modality to better estimate its class label. We introduce our approach to joint semantic segmentation of 2D & 3D…In particular, we consider a scenario where…
* The ultimate goal is to find the most probable class label for the pixels in the images and 3D points in the point cloud data

In this section, we explain our model which is defined jointly over the 2D

and 3D domains

* The entire set of 2D superpixels and 3D segments are accounted for in one graph. Let x be the set of feature computed at the nodes in the graph and y be the class labels of the nodes. We define the joint distribution over the labels given the features as: P(**y**\_2d,**y**\_3d|**x**\_2d,**x**\_3d). This probability distribution consists of different potentials detailed below.
* 2D Unary potential: This potential indicates the cost of assigning label y to the jth (i-th in our case) superpixel in the ith image, given its features, xij. The potential function Φ2D is computed as the negative logarithm of the class probabilities obtained by an SVM classifier.
* 2D Pairwise Potential: This potential function is defined over all pairwise edges which exist between the adjacent superpixels (j, k). **The potential function Ψ2D is defined in a way that penalizes dissimilar class labels for two adjacent superpixels if their RGB values are very close!!**
* This potential equals zero (via the delta indicator function) if the pair of superpixels have identical class labels.
* 2D-3D pairwise potential: This potential is applied to the edges that connect 2D

nodes to 3D nodes

* The pairwise potential Ψ2D−3D takes the relationships between the 2D objects and their 3D counterparts into account by considering pairwise links between them
* Experiments: the results of the 3D semantic

labeling is lower than that of 2D semantic segmentation.  
**My case**: the results of the labeling from the abundances is very similar to the one from the probabilities, because of their relatively high similarity as features (see the complementarity plots), but still able to complement each other.

* Table 3 and Table 4 evidence the fact that the performance of the system in each domain has been improved by incorporating information from the other domain via

2D-3D pairwise links.

* In particular, our model has increased the classification rate of the classes which hada small number of training samples, by exploiting the 2D-3D multi-correspondence.
* Note that 2D-3D pairwise edges can add inter-domain semantic information //while

they do not yield over-smoothing.

**Robust Higher Order Potential for Label Consistency longer– Pushemeet Kohli**

* A conditional random field (CRF) may be viewed as an MRF globally conditioned on the data (Lafferty et al. 2001). In this case, the potential functions are conditioned by the data and are thus should be written as ψc(xc|D). To be concise, we will drop D, and just use ψc(xc) to denote the potential functions of a CRF.
* The unary potential ψi of the CRF is defined as the negative log of the likelihood of a label being assigned to pixel i. **However, colour (unaries) alone is not a very discriminative feature and fails to produce accurate** segmentations/**classification.** This problem can be **overcome by using sophisticated potential functions.**The pairwise terms ψij of the CRF take the form of a contrast sensitive Potts model: //contrast between alphas and probs where the function g(i, j) is an edge feature based on the difference in colors of neighboring pixels. where Ii and Ij are the colour vectors of pixel i and j respectively. θp, θv, and θβ are model parameters whose values are learned using training data.
* **Inferring the Most Probable Segmentation** The object segmentation problem can be solved by finding the least energy configuration of the CRF defined above. As the pairwise potentials of the energy function (4) are of the form of a Potts model it can be minimized approximately using the well known α-expansion algorithm.

**3D Neighborhoods Point Clouds ISPRS paper**

* Conditional Random Field (CRF) for classification. CRFs are undirected graphical models that allow to model interactions between neighboring objects to be classified, and, thus, to model local context, where the objects here are the observations – alpha & probs.
* In our case, similarly to (Niemeyer et al., 2014), the nodes ni 2 n correspond to the 3D points Xi of the point cloud, whereas the edges eij 2 e connect neighboring pairs of nodes (ni, nj ). Consequently, the number of nodes in the graph is identical to the number NP of points to be classified. It is the goal of classification to assign a class label ci 2\_c1; : : : ; cL to each 3D point Xi (and thus to each node ni of the graph), where L is the number of classes, superscripts indicate specific class labels corresponding to an object type, and subscripts indicate the class label of a given point.
* we want to determine the configuration of class labels that maximizes the posterior probability p(Cjx) (Kumar and Hebert, 2006):  
  Here, Z(x) is a normalization constant called the partition function. As it does not depend on the class labels, it can be neglected in classification **?????**
* functions are called association ψ (x; ci) potentials; they provide local links between the data and the x local class labels
* functions ψ(x; ci; cj), referred to as interaction potentials, are responsible for the local context model, providing the links between the class labels (ci; cj) of the pair of nodes connected by the edge eij and the data x.
* **Association/Unary Potentials**: Note that the data appear without an index in the argument list, **x** which means that the association potential for node may depend on all the data.  
  **In my case in the unary potentials we have only dependency on i-th node/pixel data.**

The association/unary potential can be defined as the posterior probability of a local discriminative classifier based on fi(x) (Kumar and Hebert, 2006)/the features/observed data.

* **Interaction Potentials:**Consequently, we use a simpler model, namely a variant of the contrast-sensitive Potts model (Boykov and Jolly,2001) for the interaction potentials…. **Where dij (x)2 = kfi (x) 􀀀 fj (x)k2 is the square of the Euclidean distance between the node feature vectors fi (x)**

**and fj (x) of the two nodes connected by the edge eij** . Furthermore, delta\_cicj represents the Kronecker delta returning 1 if the class labels ci and cj are identical and 0 otherwise.  
//be carefull with this Kronecker delta above, this above is the opposite from mine.

The **parameter sigma is the average square distance between the feature vectors at neigh boring training points**.

* **Training and Inference**In order to determine the parameters of our classifier, we need training data, i.e. a set of points with known class labels.  
  **In case of the** association/**unary potentials**, **this involves the training of a Random Forest classifier**, **where we randomly select an identical number NS of training samples per class.  
  For the interaction potentials**, the parameter **sigma is determined as the average square distance between neighboring points in the training data based the same local**

**neighborhood** that is used for the definition of the graph in classification. For inference, i.e. for the determination of the label configuration C maximizing the posterior in Equation 5 once the parameters of the potentials are known, we use Loopy Belief Propagation (Frey and MacKay, 1998), a standard optimization technique for graphs

with cycles.

**Przemek CRF Presentation**

* **Grid search / empirical approach**  
  Applicable when number of parameters is small and their domain is known a priori better than (-Inf, +Inf) (e.g. [0;1])
  + **Empirical Approach**The weight parameters w1 and w2 (beta and gamma in my case). Here, they are set to values that were found empirically

The MRF-based model represents local behavior among the pixels in the form of the Gibbs distribution, where the value of the fused pixel is a function of pixels in the spatial neighborhood of the corresponding location. The use of the MRF-based prior typically produces spatially smooth images by favoring the radiometrically similar pixels, and penalizing sharp discontinuities.

**Phd Thesis – Wegner**

CRF

* It is a discriminative model,
* the conditional independence assumption is relaxed
* global observations may be incorporated in the association potential,
* the interaction potential is a function of both: labels and observations,
* labels and observations in the interaction potential are not limited to the local neighbourhood, but may be regarded globally.
* they directly model the posterior distribution P (yjx) of the labels y given data x as a Gibbs distribution, which leads to a relaxation of the conditional independence assumption with respect to MRFs
* This results in a much higher flexibility in terms of context formulation compared to MRFs, which are generative methods to model the joint probability P (x; y) via a distribution (cf. 2.2.3). It implies that features can be computed in spatially overlapping units in contrast to MRFs
* In addition, CRFs are globally conditioned on all data. For instance, label yd of node d is not only connected to its own data xd, but also to the data of all other nodes xa, xb, xd, and xe
* Association potential Ai (x; yi). It measures how likely a node i is labeled with yi given all data x. Iij (x; yi; yj) is also referred to as the interaction potential and it defines how the labels of two nodes i and j interact
* In particular the interaction potential Iij (x; yi; yj) is not only a function of adjacent labels yi and yj in the local neighbourhood (like in case of MRFs, compare Eq. 2.2.3), but of all data x, too.
* Improvements could be achieved by introducing, first, more discriminative features, and, second, by re-designing the unary term (global conditioning on all observed data opposed to conditioning on observed data from a specific node I where each output variable y\_i is connected with it’s observed counterpart x\_i)//the prior energy term.
* Conditional Random Fields in particular provide great flexibility for contextual classification with a single comprehensive probabilistic framework.
* Big semantically annotated training datasets should be established to unleash the full power of learning techniques.

**Semi-Supervised Discriminative Random Field for Hyperspectral Image Classification - Bioucas Dias**

At this point, we note that the posterior (1) is a particular case of a discriminative random field (DRF) [25] with association potentials given by log p(Yilxi, w) and interaction potentials given by fL b(Yi -Yj)· The DRF is based on the concept of conditional random field [26]. In a sense, it is a generalization of the Markov random field (MRF) offering several advantages, namely: i) the relaxation of conditional independence of the observed data, ii) the exploitation of probabilistic discriminative models instead of generative MRFs, and iii) the simultaneous estimation of all DRF parameters from the training data unlike the MRF framework, where the class parameters are usually estimated independently from the field parameters. In our case, we are mainly exploiting properties ii) and iii). Concerning property i), the association potential log p(Yi lXi, w) -which is strongly linked with the conditional independence- yields an excellent balance between model complexity and the quality of the results.

To apply MRF models to actual problems in low-level vision, we compute a solution using tools from probabilistic inference. Inference in this context typically means either performing maximum a-posteriori (MAP) estimation, or computing expectations over the solution space. Common to all MRF models in low-level vision is the fact that inference is challenging, both algorithmically and computationally. The loopy structure of the underlying graph makes exact inference NP-hard in the general case, although special cases exist where polynomial time algorithms are known. Because of that, inference is usually performed in an approximate fashion, for which there are a wealth of different techniques.

Classical techniques include Gibbs sampling (Geman and Geman, 1984), deterministic annealing (Hofmann et al., 1998), and iterated conditional modes (Besag, 1986). More recently, algorithms based on graph cuts (Kolmogorov and Zabih, 2004) have become very popular for MAP inference. Variational techniques and related ones, such as belief propagation (Yedidia et al.,

2003), have also enjoyed enormous popularity, both for MAP inference and computing marginals. Nevertheless, even with such modern approximate techniques, inference

can be quite slow, which has prompted the development of models that simplify inference (Felzenszwalb and Huttenlocher, 2004). While these may make inference easier, they typically give the answer to the wrong problem, as the model does not capture the relevant statistics well (cf. Fig. 2).

Inference in high-order MRF models is particularly demanding, because the larger size of the cliques complicates the (approximate) inference process. Because of that, we rely on very simple approximate inference schemes using the conjugate gradient method. Nevertheless, the applicability of more sophisticated inference techniques to models such as the one proposed here, promises to be a fruitful area for future work  
  
2.4 Other regularization methods

It is worth noting that prior models of spatial structure are also often formulated as energy terms (e. g., logprobability) and used in non-probabilistic regularization methods  
  
Volpi paper on Semantic Segmentation  
  
The terms 'i(yi) and \_ij(yi; yj) are respectively the unary and pairwise potentials, both depending on node labels yi, yj , on the parameter vectors w', w\_ learned by SSVM and on some image evidence xi, xj (the latter dropped from both potentials in Eq. (1) for clarity purposes).

This inference problem can be solved with standard graph-cuts solvers when the energy is submodular [4] or by QPBO graph-cuts when it is not [14].  
  
3.2. Pairwise potentials

Pairwise potentials are designed **to encode our prior belief about relationships between random variables**. The most commonly employed prior is the Potts smoothing, which encourages adjacent nodes to share the same label by \_p ij(yi; yj) = Jyi 6= yjK, where Jyi 6= yjK returns 1

when yi 6= yj and 0 otherwise. This potential can be made contrast-sensitive by including a term adapting to the appearance of the connected nodes. The contrast sensitive

Potts potential is defined as \_s ij(yi; yj) = g(xi; xj)Jyi 6= yjK, where **g is a function estimating the similarity of the superpixels** i and j based on the appearance descriptors xi and xj . In this case, the pairwise term encourages a label switch if the two superpixels are different under g.

The form of a standard co-occurrence pairwise potentials is \_cij(yi; yj) = h(yi; yj)Jyi 6= yjK [24]. The function h estimates a preference score between yi and yj . These potentials can be learned by counting label occurrencies in the training data and will encourage outputs with common

class associations, while discouraging rare class cooccurrences. Co-occurrence potentials can be combined with the contrast-sensitive Potts potential to account for appearance differences. Dealing with remote sensing data, co-occurrence interactions have to be estimated locally for each node. As mentioned in the introduction, most of the classes are likely to co-occur uniformly at the image level, but they cannot be learned directly from first order neighborhoods because of sparseness and bias in the ground truth.

**CRF Presentation** (Skipchain folder)

* A problem with joint distributions is that they need to explicitly model the probability of x, which can be quite complex (e.g. a textual document)
* Generative models:  
  Directed graphical models are called generative when the joint probability decouples as p(x; y) = p(xjy)p(y). The dependencies between input and output are only from the latter to the former: the output generates the input. Naive Bayes classifiers and Hidden Markov Models are both generative models
* Discriminative models:
  + If the purpose is choosing the most probable configuration for the output variables, we can directly model the conditional probability of the output given the input: p(yjx)
  + The parameters of such distribution have higher freedom  
    wrt those of the full p(x; y), as p(x) is not modelled
  + This allows to effectively exploit the structure of x without modelling the interactions between its parts, but only those with the output
  + Such models are called discriminative as they aim at modeling the discrimination between different outputs
  + //Has a short part how to compute the partition function Z

**Decision Fusion With Multiple Spatial Supports by Conditional Random Fields – Davis**

* Framework based on CRFs, where classification results obtained at pixel and region levels are probabilistically fused.
* The **aim** is to enforce the **final maps to be consistent** **not only in their own spatial supports** **(pixel and region)** **but also across supports**, i.e., by getting the predictions on the pixel lattice and on the set of regions to agree. To this end, we define an energy function

with three terms, 1) a data term for the individual elements in each support (support-specific nodes); 2) spatial regularization terms in a neighborhood for each of the supports (supportspecific edges); and 3) a regularization term between individual pixels and the region containing each of them (intersupports edges).

* 2L\_CRF improves the result obtained by the independent base model (either random

forests or convolutional neural networks)

* MRFs or CRFs) can be used to jointly account for different kinds of prior information

about spatial relationships and local likelihoods

* Owing to the Hammersley–Clifford theorem [33], random field models make it possible to express the posterior joint distribution of the unknown class labels given the observations as a Gibbs distribution. This leads to formulating the maximum a posteriori (MAP) criterion as the minimization of an appropriate energy function, modeling

conditional relationships across variables.

* Formalizing multiscale and multiresolution fusion within an MRF/CRF framework.
* We integrate the pixel- and region-based strategies within a multiscale approach, letting two (or more) spatial supports interact and come to a common decision
* These responses at all levels are fused probabilistically using a bipartite CRF
* 2L\_CRF performs structured prediction encoding:
  + 1) spatial structuring via contrast-sensitive pairwise potentials [58] and
  + 2) consistency between the labelings at the pixel and region lattices via an interlayer smoothness assumption.
* …used on purpose to emphasize that the proposed CRF model is aimed at fusing the two layers, resolving contradictions between their prediction, and “getting pixels and regions” to agree on the final labeling.
* CRFs represent a family of probabilistic models allowing to jointly characterize pixelwise class statistics as well as spatial dependencies between the labeling of different neighboring locations
* **CRF approach considers xi and yi as samples from two random fields**, i.e., a (generally continuous-valued**) random field** X = {xi }i∈I **of feature vectors** and a discrete-valued **random field Y** = {yi }i∈I **of class labels**.
* With these notations, the random field Y of the class labels is said to be a CRF if the following posterior Markovianity property holds: P(yi |y j , j \_= i,X) = P(yi |y j , j ∈ ∂i,X)
* The Markovian property:…This property makes it possible to express the posterior distribution as P(Y|X) ∝ exp[−U(Y|X)]. Minimizing such energy corresponds to the MAP criterion
* The key idea of the 2L\_CRF method is to benefit from both pixelwise and region-based image representations by introducing a novel model that connects two CRFs….to perform structured prediction at both levels simultaneously……and merges these two CRFs into a unique energy
* From the probabilistic graphical modeling viewpoint, this means using a bipartite graph to combine the two layers and generate a unique labeling
* In addition to the neighborhood system {∂ pi }i∈I p on the pixel lattice, a neighborhood system {*∂r j* } j∈Ir is also defined on the set of regions
* To apply the MAP decision rule over pixels and regions simultaneously, we consider the joint posterior distribution of the combined random field Y = (Yp,Yr ) of pixel and region labels, given the combined random field X = (X p,Xr ) of all feature vectors on both layers
* This joint posterior P(Y|X) can be expressed as a function of:
  + 1) two separate contributions associated with the marginal posteriors P(Yp|X p) and P(Yr |Xr ) of the two granularity layers, which are modeled as CRFs (see Assumption 2) and
  + 2) a cross-layer term that is related to the dependence between Yp and Yr and provides a prior on the desired relations between region labels and pixel labels
* Accordingly, the MAP rule is expressed as the minimization of the following energy function with respect to Y
* **Neighbourhood pairwise terms**:

This energy contribution is a pairwise term that favors spatial smoothness:

V\_s(ysi , ysj|Xs ) is defined as a contrast-sensitive potential that extends the classical Potts MRF model in order to **penalize that different classes** are predicted **for neighboring pixels** or regions **with similar feature vectors** [58]. //Fast Approximate Energy Minimization via GraphCuts.

* + **If two neighboring pixels have identical feature vectors** [hence, K(x pi , x p j ) = 1] and are predicted **in different classes**, then the **maximum penalty** is applied.
  + On the contrary**, if their feature vectors differ substantially** [thus K(x p i , x p j ) \_ 0], then assigning the two pixels to different classes is not (or very **slightly**) **penalized.**
  + Note that the same parameter λ is used in (6) to weigh both V p(・) and Vr (・). This is consistent with the idea of giving both granularity layers the same relevance in the labeling process. Nevertheless, extending (6) with different weight parameters for the pixel and region granularities is straightforward.
* **Consistency terms**:
  + This energy term is a cross-layer pairwise contribution that favors consistency between the labelings at the two granularity levels. This term is related to the joint prior P(Yp,Yr ) and provides **a measure of the dependence between the fields** Yp and Yr.
  + The rationale of this is to encode the desired agreement between pixelwise and region-based results
  + This behavior is favored in the proposed method by using a Potts-like formulation  
    //here he uses the Potts form only for incorporating the labels not their features, as I do ;)
  + Indeed, (9) contributes a penalty for each pixel for which the classes predicted at the two levels differ**. In my case I contribute a penalty for which the classes of these “cross -neighbours”differ and where they have similar features**. **????**
* The problem of the minimization of CRF energy functions such as (2) is generally a complex combinatorial problem. Nevertheless, in addition to consolidated stochastic optimization approaches such as simulated annealing [33], computationally efficient graph-theoretic algorithms have become very prominent during the past decade. They include graph cut algorithms, which make use of a min-flow/maxcut reformulation of the minimum energy problem [58], and belief propagation-type algorithms, which build on the idea of exchanging messages between neighboring elements to progressively reduce the energy [69].
* To minimize the resulting energy ￣U ( ￣Y |X￣ ) with respect to￣Y on the graph G (i.e., with respect to Yp and Yr simultaneously), the sequential tree reweighted message passing algorithm is used in 2L\_CRF.
* Experimental Setup:  
  We compare the results obtained using models aware of a single support (i.e., only pixel- or region-based mapping units) with those obtained by models exploiting structure in the spatial domain of the support (in all cases, the CRF of [51]) and with the results of the proposed model, which lets the two supports interact via the two-layers CRF structure.  
  We consider two base classifiers: RFs [64] and CNNs [21]
  + RFs: In the case of RFs, we train two separate models, one at the pixel and another at the region level [60]. At the pixel level, we use spectral features (R-G-B-NIR raw values, normalized to standard scores),…  
    To train the models, we derived a region ground truth from the available pixelwise training map by assigning the majority class found within each region.
  + CNNs:
* Finally, for both data set hyperparameters μ and λ were set by cross-validation and the σ parameter of the kernel in the contrast sensitive term (8) was set as half the Euclidean

features distance between samples.

* Experimental Results:
  + RFs:   
    The slight increases in accuracies are due to the larger training set used to train the pixel-based classifier and are consistent across the experiments

//This is in line with observations in several other recent papers.

* The 2L\_CRF model only needs to correct for inconsistent labelings among the two layers, which results in small increases in the accuracies
* On the contrary, ….allows to correct for several misclassifications at the region level
  + Parameter Analysis:
    - Fig. 7 reports the sensitivity analysis for the μ and λ parameters, involved in (6). As a reminder, μ control the strength of the relations between the spatial supports and λ the strength of the local spatial smoothing within each spatial support
    - In all experiments, the model not making any use of intersupport connections (μ = 0, which would roughly correspond to optimizing two separate CRFs) does not provide the best results, **thus showing the interest of a joint model**.
    - This is particularly visible at the region level, where the evidence from the pixel level is able to successfully revert many errors, in particular for the RF classifier, which often provides ambiguous class-likelihoods  
      **In my case**: evidence from the MLR is able to improve many errors coming from the Unmixing which (probably) provide more ambiguous class-likelihoods (the entropy for the abundances is higher) from the fact that their original nature is physical and not statistical.
    - Also, using **contrast sensitivity seems particularly important at the pixel level**, while less at the region level: **this is expected**, since given the high resolution of the data, **local inconsistencies at the pixel level are more frequent** than noisy predictions at the region level
    - **In my case for the CRF method**: more local inconsistencies in Pavia are to be expected since the image has higher resolution meaning more complexity, with initial classif. maps which might contain more artefacts/or peper and salt noise**????** and in order to account for that /leading to   
      there is a higher penalization of the neighbors->higher beta value (closer to its gamma value).   
      While for Pines there is not such high complexities due to lower resolution of the image, so the penalization is lower -> lower beta compared to it’s gamma value. The much lower beta (40 times) compared to it’s gamma comes from the fact that in order for the cross-edge term to have an effect/influence, gamma has to be much higher than the neighborhood term/beta value.
    - **Maybe this is just blab la, but I left it here for ideas for my case**:  
      therefore beta\_pavia is 2xhigher than beta\_pines. Beta\_pavia is higher since it tries to penalize more local neighborhoods in Pavia, for the reason of having higher complexity in the image, //for neighbors with different classes the neighbouring //abundances/probabilities are more alike, coming from the fast that //they are computed for smaller areas (covering the same //objects/materials). **????**

While beta\_pines is lower penalizes less local neighborhoods:

//for neighbors with different classes the neighboring abundances /probabilities are more distinct coming from the fact that they are //computed for larger areas (covering different objects), leading to //smaller penalization **????**

* + - Davis paper again: It still suffers from some prediction noise leading to small artifacts, as it can be seen in the maps in Fig. 8 for the entire tiles and in Fig. 9 for details. This is certainly due to the increased spatial resolution (and therefore complexity) of the problem, since we are working, compared to the Zurich Summer data set, on a 10 times higher spatial resolution
    - The use of CRFs with up to pairwise nonzero clique potentials will make the label swaps/improvements if the pixel classifier (MLR or Unmixing) is affected by salt and pepper noise (which is my case, see Unmixing and MLR classification maps, they have salt and pepper noise). This is corrected by looking only at the direct pixel/node neighborhood.
    - **My case MRF method beta parameters:**  
      MRF works best when all interactions receive more or less equal weights. MRF is using the same potentials (Kronecker deltas) for both the pairwise and the cross edge interactions.
  + **Conclusions:**
    - We proposed a probabilistic discriminative graphical model relying on a CRFs formulation for the fusion of land-cover and land-use classification results. The system is able to find agreement between probabilistic decisions with multiple spatial supports
    - We explored the fusion of pixel- and region-based spatial supports within an energy minimization framework, with contributions from:
      * The posterior distribution of the land cover classes at the single instance level (the pixel or the region);
      * The spatial smoothness of the predictions (i.e., the consistency of the prediction among the spatial neighbors);
      * The smoothness across supports (i.e., the consistency of the predictions between a region and the pixels composing the region itself
    - These three goals are addressed jointly within a CRF model with connections across layers corresponding to different spatial support representations.
    - Applications to two VHR benchmark data sets showed the potential of the approach that, using common models (we considered RFs and CNNs), can improve the final maps consistently and joining the spatial detail of the pixel support with the geometrical object accuracy of the region support
    - **In my case:** In the future……it would be interesting to train beta and gamma values directly from data.

**[A higher order conditional random field model for simultaneous classification of land cover and land use](D:\\Documents\\Remote Sensing\\Data Fusion\\CRF\\28.A higher order conditional random field model for simultaneous classification of land cover and land use.pdf) ISPRS 2017**

* Edges, which model statistical dependencies between the class labels and data at the associated nodes
* CRF are discriminative classifiers, thus, directly modelling the posterior probability PðyjxÞ of the label vector y given the data x
* One can choose arbitrary discriminative classifiers with a probabilistic output P(y\_i|**x**) for the association potential. For pairwise potentials, one can also apply any discriminative classifier with a probabilistic output
* In order to realize a simultaneous classification of land cover and land use, where both classification tasks mutually support each other, we design a graphical model consisting of two layers. The layers correspond to hierarchical levels and are arranged one above the other, connected by **inter-layer edges**. Each layer consists of nodes and **intra-layer edges**.
* We want to determine the class labels for land cover y\_c\_i and land use y\_u\_k

in the corresponding layers. **In my case we want to determine the labels from source 1:y\_alpha or source 2:y\_p, and use as a result the labels irrespectively from the source, having as primary goal the improvement of the classification using low training data. // either from source 1 or 2**

* Both layers differ with respect to the image entities represented by the nodes and the employed features.
* The classes: are y\_i\_c belongs to L\_c and y\_k\_u belongs to L\_u
* The intra-layer edges model the spatial neighbourhood of each node in the respective layer. The neighbourhood of a node ni is composed of its first-order spatial neighbours, i.
* Inter-level context, i.e. the statistical dependencies of land cover and land use, are modelled via inter-layer edges
* **Pairwise term 1 and pairwise term 2** represent the pairwise intra-layer interaction potentials, which model the spatial dependencies between neighboring sites within each layer in consideration of the data x.  
  **Consistency term** - describes the inter-layer higher-order potential, which models the relations between the labels yc in the land cover layer and yu in the land use layer
* The parameters omega = {w1,w2,…w5} determine the impact of each potential relative

to the first potential term

* **Association potential**: The data are taken into account in the form of site-wise feature vectors f\_c\_i(**x**) and f\_k\_u(**x**) for the nodes in the land cover and land use layer, respectively. Both association potentials take values proportional to the probability of yci and yuk given the site-wise feature vectors f\_c\_i(**x**) and f\_k\_u(**x**),   
  i.e. ksi(y\_c,**x**) ∝ P(Y| f\_c\_i(**x**))…**see the rest from the paper**
* **Pairwise intra-layer potential**This potential models the dependencies of the labels of nodes ni and nj being adjacent within one layer, considering the data x. The probability of a certain class relation is learned from real-world occurrences in representative training data. Thus, the interaction potential is modelled as the joint posterior probability of both labels yci and ycj given their data, i.e. **see paper!!!!**
* **Inter-layer higher order potential**This potential models the semantic dependencies between land cover labels yci and land use labels yuk of all nodes nci and nuk belonging to a higher order clique h/the whole graph in my case, which is formed by all image sites in both layers (having a spatial overlap). The inter-layer potential takes a value proportional to the joint probability of the set of class labels ych in the land cover and yuh in the land use layer of all nodes nh being connected in a higher order clique h, i.e. **see paper!** The data are not taken into account for the estimation of this potential. Thus, the inter-layer potential only depends on class labels and their respective beliefs. **In my case** the observed thata: alpha\_i and p\_i are takin into account for the estimation of this potential, so the inter – consistency potential depends on the class labels and the observed data. **In my case**: the potential should favour all land cover sites to take the same label.
* Hence, learning all possible class configurations by a statistical classifier becomes computationally intractable. Furthermore, by involving a large number of variables in a generic formulation of a higher order potential, inference becomes intractable as well.
* In the inference step, the most probable label configuration y is determined for all nodes in a CRF simultaneously. This is based on maximizing the posterior probability PðyjxÞ of the labels given the data. Exact inference is computationally intractable for multi-class problems (Kumar and Hebert, 2006). Therefore, only approximate methods can be used.  
  An approximate solutions: Loopy Belief Propagation (LBP)
* The **optimal parameters** for the RF classifiers applied for each potential term as well as for the **inference procedure were determined empirically**
* From all available training samples in a test run, we randomly select Nt samples per class (cf. Table 2) in order to avoid a bias for classes for which a large number of samples is available.
* The **classification result** to the reference for **each layer separately!!!!**
* We want to investigate the impact of the extracted features on the classification result in order to determine an appropriate set of features. Therefore, the relevance of each feature in the classification process is analyzed based on**…. In our case** we don’t investigate the impact of the extracted features, since the focus is on the fusion frameworks, but I could add this as an improvement!
* **I could also have a parameter settings table!**
* The results are not compared to the results of the two-step processing strategy**, because the land cover classification is not improved during the two-step procedure** (used as input for the land use classification). **In my case** I could say: that the second classification map has improvements as well (or both classification maps are improved), with some slight differences, but since our primary goal is general improvement of the classification using low # of train data with the fusion framework we present the results from the improved classification from one of the sources, just to illustrate that there is an improvement, irrespectively from the resulting source.

**On Learning Higher-Order Consistency Potentials for Multi-class Pixel Labeling**

* where [[・]] is the indicator function which takes 1 when the argument is true and 0 otherwise, xi is the color vector for pixel i, and λ and β are global and image specific constants that determine the strength of the smoothness prior.
* E. Note that the contrast sensitive smoothness prior (Eq. (2)) is submodular (they talk about binary variables here!!!).
* The goal of inference is to find the assignment y\_kappa with minimum energy. Message-passing algorithms can be suitable for the objective, but it is well known that for submodular pairwise energy functions this can be done efficiently by finding the minimum-cut in a suitably constructed graph [14,15,16]. Unfortunately, in general for multi-label CRFs (or indeed, non-submodular binary CRFs), inference is intractable and we need to resort to approximate routines
* **Move-Making Inference**. Generally, most energy minimization problems are NP-hard [13]. However, there are good approximate solutions available. For example, some move-making algorithms reduce the energy minimization to sequence of smaller problems which are submodular. Here, each move restricts the label space of variables to at most two values from the label set.
* An early example of move-making algorithms is Iterated Conditional Modes (ICM) [17]. Its convergence slow and can easily get stuck in poor local optima. The more advanced examples of move-making algorithms are α-expansion and αβ-swap [15]. In α-expansion, a label from L is chosen iteratively. Each variable can switch to the chosen label α or keep the current label, which expands the current label α to other regions as long as the energy is reduced
* In summary, the three algorithms are characterized as follows:
  + ICM
  + α-expansion
  + αβ-swap
* For α-expansion and αβ-swap, an efficient graph-cut based algorithm has been proposed to minimize the energy functions composed of pairwise potential functions

[Building Detection From One Orthophoto and High-Resolution InSAR Data Using Conditional Random Fields](file:///D:\Documents\Remote%20Sensing\Data%20Fusion\CRF\Building%20Detection%20From%20One%20Orthophoto%20and%20High-Resolution%20InSAR%20Data%20Using%20CRFs.pdf) **(Wegner)**

* One drawback of MRFs is that they generatively model the joint distribution of labels and data. This implies that the distribution of the data has to be modeled which is often hard to accomplish. Conditional Random Fields (CRF) are discriminative models and hence the data distribution does not have to be modeled
* In this paper, for the first time, a CRF is used for combining optical and SAR data
* **In my case**: for the first time a CRF is used for combining unmixing and classifier data**????**
* An approach for building detection in an urban area that combines features from an orthophoto with line features from mono-aspect InSAR data. Both feature sets are introduced to a single feature vector, used to distinguish 2 classes in a CRF framework.
* We compare the CRF to two other probabilistic methods: a standard Maximum Likelihood classifier (non-contextual) and a state-of-the-art Markov Random Field (contextual).
* MRFs can be seen as an extension of Naïve Bayes. They are generative models and hence estimate the joint distribution of data and labels P(**x**,**y**) , which can be decomposed into a product of factors P(**x**|**y**)P(**y**).
* The likelihood uses data only from a single site and not from all sites (like CRFs). The prior term only compares adjacent labels y\_j with the investigated label y\_i .
* In contrast, **CRFs** are based on the maximum entropy approach, which is known to be able to **provide accurate and robust classification results**. As opposed to MRFs, CRFs are discriminative models and therefore model only the posterior distribution P(**y**|**x**) of the labels given data .
* The association potential measures how likely a site i is labeled with y\_i given the data **x** , while the interaction potential describes how two sites i and j interact.
* It should be noted that both potentials have access to the whole image. In particular the interaction potential is not only a function of adjacent labels y\_i and y\_j in the local neighborhood (like in case of MRFs, compare (1)), but of all data **x**, too.
* Association Potential: A\_i(**x,**y\_i) = exp(y\_i\*w’**h\_i**(**x**)), while   
  P(**y**|**x**) = 1/Z(**x**) \*exp(sum\_i{A\_i(**x,**y\_i)} + sum{sum{I\_ij(**x**,y\_i,y\_j**)}}), so they don’t use log of the features: h\_i(x)!!!!!**
* The interaction potential determines how two sites i and j should interact regarding all data **x.** y\_i is the label of the site of interest and y\_j the label it is compared to

**Minimizing non-submodular functions with Graphc Cuts PAMI07-QPBO**

* Many early vision problems can be naturally formulated in terms of energy minimization

where the energy function has the following form: …  
This energy is often derived in the context of Markov Random Fields [6,15]: a minimum of E corresponds to a maximum a-posteriori (MAP) labeling x.

* Indeed, one of the most successful MRF minimization algorithms, namely the expansion move method of Boykov et al. [12], reduces the problem with multi-valued variables to a sequence of minimization subproblems with binary variables.
* There is an accepted view within the computer vision community popularized by [24] that graph cuts can only be used for minimizing submodular energy functions, i.e. functions whose pairwise terms satisfy \_pq(0, 0) + \_pq(1, 1) \_ \_pq(0, 1) + \_pq(1, 0) .
* For functions of multi-valued variables and the expansion move algorithm the corresponding condition is \_pq(\_, )+\_pq(\_, \_) \_ \_pq(\_, \_)+\_pq(\_, ) which must hold for all labels \_, \_, 2 {0, . . . ,K − 1} 2. While many important energy functions in vision (e.g. Potts) do satisfy these conditions, in some situations we get functions which are not submodular. For example, they may arise when parameters of the energy function are learned from training data. In my case the parameters of the energy functions are: beta and gamma, they are not learned from the data, but are empirically set, but the sigmas inside the pairwise CRF potentials are learned from training data. Does this make my energy function non submodular? I thought that when we use a Potts pairwise potentials we always end up with a submodular function**??? See below about metric and semi-metric interaction potentials!! (My interaction potential has to be metric in order to be able to use the alpha expansion method). Davis uses the (extension) to Potts model (Fast Approximate Energy Minimization via GraphCuts) in his interaction potential for the CRF as I do!!  
  He also has parameters trained from data inside the energy function: lambda and gamma he gets it by doing cross validation and sigma in the Gaussian Kerne he sets it as half the Eucledian distance between samples (so he learns all parameters from train data)l!!!   
  And nowhere he says that his function is not submodular!!!**
* It is therefore desirable to have a method which explicitly takes into account non-submodular terms instead of throwing them away. In this paper we will review the algorithm in [10,19] for minimizing functions with both submodular and non-submodular terms. We refer to it as the QPBO method (“quadratic pseudo-boolean optimization”)

[Fast approximate energy minimization via graph cuts](file:///D:\Documents\Remote%20Sensing\Data%20Fusion\GraphCuts\Fast%20approximate%20energy%20minimization%20with%20Graph%20Cuts.pdf)  
  
//expansion move method which reduces the problem with multi-valued variables to a //sequence of minimization subproblems with binary variables

* The major di\_culty with energy minimization for early vision lies in the enormous computational costs. Typically these energy functions have many local minima (i.e., they are non-convex). Worse still, the space of possible labelings has dimension jPj, which is many thousands.
* Simulated annealing was popularized in computer vision is widely used since it can optimize an arbitrary energy function. Unfortunately, minimizing an arbitrary energy function requires exponential time, and as a consequence simulated annealing is very slow. In practice, annealing is ine\_cient partly because at each step it changes the value of a single pixel.
* In special cases such energies can be minimized exactly. If the number of possible labels is jLj = 2 then the exact solution can be found in polinomial time by comput-ing a minimum cost cut on a certain graph [4]. If L is a \_nite 1D set and the interaction potential is V (fp; fq) = |fp−fq| then the exact minimum can also be found e\_ciently via graph cuts [5, 2]. In general, however, the problem is NP-hard.
* In this paper we develop algorithms that approximately minimize energy E(f) for an arbitrary finite set of labels L under two fairly general classes of interaction potentials V : semi-metric and metric.
* V is called a **semi-metric** on the space of labels L if for any pair of labels alpha, beta belonging to L it satisfies two properties:
  + V (alpha,beta) = V (beta,alpha) >= 0 and V (alpha,beta) = 0 , alpha = beta.  
    //**In my case the interaction potential is at least semi-metric!!!**

If V also satisfies the triangle inequality:

* + V (alpha,beta) <= V (alpha,gamma) + V (gamma, beta)

for any alpha, beta, gamma in L then V is called a **metric**.

* Note that both semi-metric and metric include important cases of discontinuity-preserving interaction potentials. For example, **the Potts interaction penalty:** V = delta(y\_i != y\_j) **is metric.**
* The algorithms described in this paper generalize the approach that we originally developed for the case of the Potts model
* Our second algorithm, described in section 4, is based on more interesting alpha-expansion moves but works only for metric V(y\_i,u\_j) interaction potentials (i.e., the additional triangle inequality constraint is required). Note that alpha-expansion moves produce a solution within a known factor of the global minimum of E. A proof of this can be found in [8].
* Experimental Results:  
  For our experiments, we used three energy func tions, each with a quadratic Dp. The first energy function, called E1, uses the truncated quadratic   
  **V\_i,\_j(****y\_i; y\_j) =** **min(K; |y\_i – y\_j|^2)** (for some constant K) as its smoothness term. **This choice of V does not obey the triangle inequality, so we minimized E1 using our swap move method.** **The second (E2) and the third (E3) energy functions use, correspondingly, the Potts model** and the truncated L2 distance as their smoothness penalty V . **Both of these obey the triangle inequality and we minimized E2 and E3 with our expansion move method**.

**MAP/Max Likelyhood**MAP/Max Likelihood (most probable class from the classifier for each pixel independently) or prediction of independent class labels y\_i.

**Motivation for using CRF over MRF:**See Davis presentation:

* MRF - independence assumption: conditional independence of the observed data, the observed pixels x\_i are independent knowing the class y: P(**x**|**y**) = P(x1|y)\* P(x2|y)\*… P(xn|y). //Bioucas paper above.
* MRF uses only the labels as random fields & relationships between the labels in the pairwise potential
* **CRF uses both: the labels and the observed data as random fields in the pairwise potential and thereby includes relationships/similarities not only between labels but also between observed data**, since we would like to make discrimination between neighbors that look alike and have/belong to different classes.
* CRF: The observed data points: x\_1, x\_2, … x\_n given the class y are not assumed to be independent anymore as in the MRF (relaxing the conditional independence assumption of the observed data). In reality the observed data are correlated which is also the case with our observed data: the neighboring abundances and the probabilities, these (for a known class) are not independent of each other.
* MRF produces smooth boundaries, while CRF sharper boundaries  
  MRF issue: at the corners the neighboring pixels/nodes will have different classes, the unary potential from MRF produces one class while the neighbors have another class, so the MRF tries to be consistent with the neighbours, therefore it changes it’s class to the neighbors one, making a mistake & producing smooth region. Therefore we need not only consistency with the neighbor labels but also consistency with the observed data from the neighbours. In order to accomplish this we add the exp(-||a-b||) term penalizing neigbouring labels with different classes having similar observed data.
* **Daphne Coller** video:

These features are more informative about the class of the pixel i and are very correlated with each other. So if we have ve ry correlated features containing lots of redundant information, if we represent this in a very naïve bayes model, where the features should be independent given the label // P(X1|Y),P(X2|Y),…P(xn|Y) we are ignoring that correlation structure. They make incorrect independence assumption. In standard modeling, we model the joint distribution P(X,Y) //probability of X and Y together. Instead of doing this, we will model, a conditional distribution of Y given X: P(Y|X), where we are NOT trying to capture the distribution over X (the features). So if we are not trying to capture the distribution over X we don’t care about the correlation of the features. We can do the modeling of P(Y|X) with the Conditional Random Fields – CRF. **In our standard modeling for P(X,Y) we only have pairwise terms which relate the Xi with Y**, **we don’t have any terms that relate X1, X2,…Xn (the features) to each other.** With the conditional distribution modeling: P(Y|X) we remove from the analysis every notion of a correlation between the X1,X2,..Xn features. We are just modeling how the X1,X2,…Xn come together to affect the probability of Y. So with these CRF models we have the ability to ignore the distribution over the features and focus only on the target variables, and allow us to sort of ignore correlations between features and not worry about whether they are independent of each other or not.

* Experiments and Discussion

--------------------------------------------------  
Write something similar as this:   
We randomly select few labeled samples from the groundtruth map to generate the training set as well as the dictionary, and the remaining samples are used for testing. Let Ll and Lu be the number of initial training set and unlabeled samples, respectively. In order to illustrate the generalization performance of the proposed methods, we use very few labeled training sets on purpose. For performance comparison, we compare our methods with some state-of-the-art classifiers listed as follows.  
from: Sparse Graph Regularization for Hyperspectral Remote Sensing Image classification paper  
--------------------------------------------------

* + **Sensitivity analysis**:  
    **beta & gamma**:  
      
    Pavia:
    - MRFL: beta = 0.6/(2\*4) = 0.075, gamma = 0.6,   
      beta/gamma = 0.125 = 12.5%
    - CRFL: beta = 10/(2\*4) = 1.25, gamma = 5  
      beta/gamma = 1.25/5 = 0.25 = 25%

Pines:

* + - MRFL: beta = 0.8/(2\*4) = 0.1, gamma = 0.9  
      beta/gamma = 0.111 = 11%
    - CRFL : beta = 5/(2\*4) = 0.62, gamma = 25  
      beta/gamma = 0.124 = 12.4%  
        
      Having nonzero values for the beta and gamma parameters tells us that the assumption of neighborhood label consistency & across label consistency holds & agrees with our prior belief.  
        
      Penalization happens for neighbors with different labels for MRF and different labels and similar observations for CRF:

From the MRFL:   
Pavia: beta (per pairwise interaction) is 12.5% smaller than the gamma   
Pines: beta (per pairwise interaction) is 11% smaller than gamma  
Conclusion: the influence of the gamma/consistency term is about 12% higher than the neighborhood term, since the output labels from source 1 and source 2 differ more than labels within the same source so it tries to penalize higher differences between output labels from different sources on top of the initial penalization happening inside the pairwise potentials.

From the CRFL:   
Pavia: beta (per pairwise interaction) is 25% lower than gamma  
Pines: beta (per pairwise interaction) is 12.4% lower than gamma  
Here again we can see that the influence of the gamma parameter is higher than beta. The same situation as in the MRF case with the difference that now in the pairwise potential the observed abundances and probs are used on top of the labels.   
Having higher gamma is the result of trying to additionally penalize higher differences between abundances from source 1 and it’s corresponding prob from source 2, since the abundances are more dissimilar to the probabilities on top of the initial penalization happening inside the pairwise function, while the penalization is lower (beta is lower) for neighborhood terms within the source 1 (within abundances themselves) and within source 2 (within probabilities themselves) since the difference within abundances and within probabilities is lower – they are more alike, they were already highly penalized inside the pairwise potential function.  
  
For Pavia the beta influence is for 25% lower than gamma compared to 12.4% in Pines, so the neighborhood terms influence more in the case of Pavia compared to Pines. I still don’t understand why and how is this related to the spatial resolution. See my nice drawing ☺

It’s noticeable that the classification accuracies can be improved when the values for β and gamma are increased. This can certify the effectiveness by incorporating both terms, the spatial neighborhood term and the consistency term in our proposed methods.//leads to improvement of the overall classification.  
//the accuracy is more sensitive to changes in the beta parameter value, while for different //gamma values the accuracy changes as well in a subtle way.  
  
**lambda**: //read collaborative… filtering from Iordache to see the parameter values  
Add: lambda vs. fusion accuracy (MRFL) for Pavia & Pines to show   
that less sparse abundances are more useful for the overall fusion process **?????**

* **Effect on using different features for pairwise fusion**

P\_alpha - these features have their own advantages and disadvantages

* **Effect on using additional source for fusion**
* **Results from different classification methods**

The proposed fusion method produces in general higher accuracies (overall and average) compared to the other methods.

//Description about the classification maps:  
For the reason of U and MLR being classifiers without incorporating any spatial information and where no fusion is performed, they produce noisy classification maps, while the MRF\_a, MRP\_p, CRF\_a, CRF\_p on the other side they do incorporate spatial information and produce therefore less noisy classification maps //due to the combination of both spectral and spatial information. In contrast, our fusion methods manage to……

To illustrate the superiority of the proposed methods, we conduct pairwise McNemar statistical test between classifiers.   
In the 11 pairwise tests over both scenes – Pines and Pavia, in 10 our methods achieved a better significant result than the other methods. //check this 11 and 10 numbers!!!   
Should I add this**??????   
What would be the reason that MRFL or CRFL for Indian Pines doesn’t give me more significant classification results than the others?**

Thesis: Graph-based Data Modeling and Analysis for Data

Fusion in Remote Sensing  
  
In the MRF framework, the maximum a posteriori (MAP) decision rule is typically formulated as

an iterative optimization step of the energy function, which is extremely time consuming

with high resolution data. Furthermore, classical models used in MRF framework

(e.g., Ising, Potts model) suffer from the high spatial resolution: neighboring pixels are

highly correlated, while the standard neighbor system definition does not contain enough

samples to be effective.  
  
Morphological profiles  
  
The other approach comprises multi-scale techniques as well as an adaptive neighborhood

of a pixel according to the structures to which it belongs and was proposed by

Benediktsson, et al [25]. They have exploited the morphological filters as an alternative

way of performing joint spatial-spectral classification. Rather than defining a crisp

neighbor set for each pixel, morphological filters enable the definition of an adaptive

neighborhood of a pixel according to the size and shape of the structures to which it

belongs. This adaptive neighborhood approach has shown its good performance for

multispectral and hyperspectral data [26]. Recently, M. Dalla Mura et al. [27, 28] investigated

the connected morphological operators for the analysis of very high resolution

images, as well for HSI as an extension of the morphological profile based on a series of

attribute filters.  
  
Future work:  
  
CRFs have the advantage of incorporating global context information, while in this work we incorporated only local neighbourhood/contect information from observed data.  
The likelihood in MRF uses data only from a single site and not from all sites (like CRFs). // see papers from Wegner as: Building Detection From One Orthophoto and

High-Resolution InSAR Data Using Conditional

Random Fields and his Phd Thesis.