HIERARCHICAL CONDITIONAL RANDOM FIELD FOR MULTI-CLASS IMAGE CLASSIFICATION

Michael Ying Yang, Wolfgang Förstner
Department of Photogrammetry, Bonn University, Bonn, Germany
michaelyangying@uni-bonn.de, wf@ipb.uni-bonn.de

Martin Drauschke
Institute for Applied Computer Science, Bundeswehr University Munich, Munich, Germany
martin.drauschke@unibw.de

Keywords: Multi-class image classification, hierarchical conditional random field, image segmentation, region adjacency graph, region hierarchy graph.

Abstract: Multi-class image classification has made significant advances in recent years through the combination of local and global features. This paper proposes a novel approach called hierarchical conditional random field (HCRF) that explicitly models region adjacency graph and region hierarchy graph structure of an image. This allows to set up a joint and hierarchical model of local and global discriminative methods that augments conditional random field to a multi-layer model. Region hierarchy graph is based on a multi-scale watershed segmentation.

1 INTRODUCTION

In recent years an increasingly popular way to solve various image labeling problems like object segmentation, stereo and single view reconstruction is to formulate them using image regions obtained from unsupervised segmentation algorithms. These methods are inspired from the observation that pixels constituting a particular region often have the same label. For instance, they may belong to the same object or may have the same surface orientation. This approach has the benefit that higher order features based on all the pixels constituting the region can be computed and used for classification. Further, it is also much faster as inference now only needs to be performed over a small number of regions rather than all the pixels in the image.

Classification of image regions in meaningful categories is a challenging task due to the ambiguities inherent to visual data. On the other hand, image data exhibit strong contextual dependencies in the form of spatial interactions among components. It has been shown that modeling these interactions is crucial to achieve good classification accuracy, (cf. Section 2).

Conditional random fields (CRFs) have been proposed as a principled approach to modeling the interactions between labels in such problems using the tools of graphical models (Lafferty et al., 2001). A conditional random field is a model that assigns a joint probability distribution over labels conditioned on the input, where the distribution respects the independence relations encoded in a graph. In general, the labels are not assumed to be independent, nor are the observations conditionally independent given the labels, as assumed in generative models such as hidden Markov models. The CRF framework has already been used to obtain promising results in a number of domains where there are interactions between labels, including tagging, parsing and information extraction in natural language processing (McCallum et al., 2003) and the modeling of spatial dependencies in image interpretation (Kumar and Hebert, 2003).

One problem with the methods using low-level features in image classification is that it is often difficult to generalize these methods to diverse image data beyond the training set. More importantly, they lack semantic image interpretation that is valuable in determining the class labeling. Contents such as the presence of people, sky, grass, etc., may be used as cues for improving the classification performance obtained by low-level features alone.

This paper presents a proposal of a CRF that simultaneously models the region adjacency graph and the region hierarchy graph structure. This allows to set up a joint and hierarchical model of local and global discriminative methods that augments CRF to a multi-layer model.

The contributions of this paper are the following. First, we extend classical one-layer CRF to multi-layer CRF while restricting to second-order cliques. Second, this work shows how to integrate local and global information in a powerful model. The paper is organized as follows: Section 2 introduces related work. Section 3 gives the basic theory of CRF. Section 4 presents pairwise CRF model by incorporating novel hierarchical pairwise potentials.

2 RELATED WORK

There are many recent works on multi-class image classification that address the combination of global and local features (He et al., 2004; Yang et al., 2007; Reynolds and Murphy, 2007; Gould et al., 2008; Toyoda and Hasegawa, 2008; Plath et al., 2009; Schnitzspan et al., 2009). They showed promising results and specifically improved performance compared to making use of only one type of features - either local or global.

He et al. (2004) proposed a multi-layer CRF to account for global consistency and due to that showed improved performance. The authors introduce a global scene potential to assert consistency of local regions. Thereby, they were able to benefit from integrating the context of a given scene. However, their model works with global priors set in advance and only uses learned local classifiers. Rather than to rely on priors alone, in our work, all parameters of the layers are trained jointly. Yang et al. (2007) proposed a model that combines appearance over large contiguous regions with spatial information and a global shape prior. The shape prior provides local context for certain types of objects (e.g., cars and airplanes), but not for regions representing general objects (e.g., animal, building, sky and grass). In contrast to this, we explicitly model hierarchical graph structure of an image, capturing long range dependencies. Gould et al. (2008) proposed a method for capturing global information from inter-class spatial relationships and encoding it as a local feature. Toyoda and Hasegawa (2008) presented a proposal of a general framework that explicitly models local and global information in a conditional random field. Their method resolves local ambiguities from a global perspective using global image information. It enables locally and globally consistent image recognition. But their model needs to train on the whole training data simultaneously to obtain the global potentials, which results in high computational time.

Besides the above approaches, there are more popular methods to solve multi-class classification problem using higher order conditional random fields (Kohli et al., 2007, 2009; Ladicky et al., 2009). Kohli et al. (2007) introduced a class of higher order clique potentials called P^n Potts model. Higher order clique potentials have the capability to model complex interactions of random variables, making them able to capture better the rich statistics of natural scenes. The higher order potential functions proposed in Kohli et al. (2009) take the form of the Robust P^n model, which is more general than the P^n Potts model. Ladicky et al. (2009) generalized Robust P^n model to P^n based hierarchical CRF model. Inference in these models can be performed efficiently using graph cut based move making algorithms. However, the work on solving higher order potentials using move making algorithms has targeted particular classes of potential functions. Developing efficient large move making for exact and approximate minimization of general higher order energy functions is a difficult problem. Parameter learning for higher order CRF is also a challenging problem.

Recent work by Plath et al. (2009) comprises two aspects for coupling local and global evidences both by constructing a tree-structured CRF on image regions on multiple scales, which largely follows the approach of Reynolds and Murphy (2007), and using global image classification information. Thereby, Plath et al. (2009) neglects direct local neighborhood dependencies, which our model learns jointly with long range dependencies. Most similar to us is the work of Schnitzspan et al. (2008) who explicitly attempt to combine the power of global feature-based approaches with the flexibility of local feature-based methods in one consistent framework. Briefly, Schnitzspan et al. (2008) extend classical one-layer CRF to multi-layer CRF by restricting pairwise potentials to 4-neighborhood model and introducing higher-order potentials between different layers. There are several important differences with respect to our work. First, rather than 4-neighborhood graph model in (Schnitzspan et al., 2008), we build region adjacency graph based on watershed image partition, which leads to a irregular graph structure. Second, we apply an irregular pyramid to represent different layers, while Schnitzspan et al. (2008) use a regular pyramid structure. Finally, our model only exploits up to second-order cliques, which makes learning and inference much easier. While Schnitzspan et al. (2008) introduce higher-order potentials to represent interactions between different layers.

3 PRELIMINARIES

We start by providing the basic notation used in the paper. Let the image $\mathbf{X}$ be given. It is described by a set of regions with indices i collected in the set $R = \{i\}$.

They are possibly overlapping and not necessarily covering the image region. Multi-class image classification is the task of assigning a class label $l_i \in \mathcal{C}$ with $\mathcal{C} = \{1, \ldots, C\}$ to each region i.

Let $G = (R, E)$ be the graph over regions where E is the set of (undirected) edges between adjacent regions. Note that, unlike standard CRF-based classification approaches that rely directly on pixels, e.g., (Shotton et al., 2006), this graph does not conform to a regular grid pattern, and, in general, each image will induce a different graph structure.

The conditional distribution of a classification for a given image has the commonly general form

$$P(\mathbf{L} \mid \mathbf{X}) = \frac{1}{Z} \exp\left(\sum_{i \in R} f_i(l_i \mid \mathbf{X}) + \sum_{(i,j) \in N} f_{ij}(l_i, l_j \mid \mathbf{X}) \right) \tag{1}$$

where $\mathbf{L} = \{l_i\}_{i \in R}$ represent the labeling of all regions, N is the set of neighbored regions , and Z is the partition function for normalization. The unary potential f_i represents relationships between labels and local image features. The pairwise potential f_{ij} represents relationships between labels of neighboring regions.

The unary potential f_i measures the support of the image $\mathbf{X}$ for label l_i of region i. Various local image features are useful to characterize the regions. For example, the CRF in (Shotton et al., 2006) uses shape-texture, color, and location features. The pairwise potential f_{ij} represents compatibility between neighboring labels given the image $\mathbf{X}$. E. g. if neighboring regions have similar image features, f_{ij} favors the same class label for them. Then, if the regions have dissimilar features, they might be assigned different class labels. Thus, the pairwise potential f_{ij} supports data-dependent smoothing.

4 HCRF: HIERARCHICAL CONDITIONAL RANDOM FIELD

While global detectors have been shown to achieve impressive results in image classification for unoccluded image scene, part-based approaches tend to be more successful in dealing with partial occlusion. Since adjacent regions in images are not independent from each other, CRF models these dependencies directly by introducing pairwise potentials.

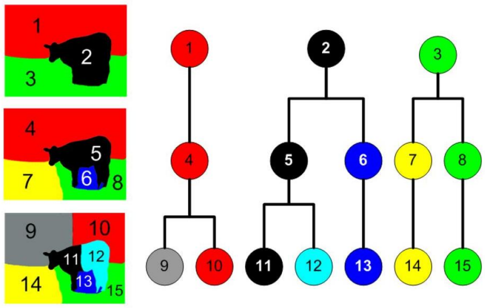

Figure 1: Simulated segmentations at three scales (left), with corresponding region hierarchical graph (right) (Reynolds and Murphy, 2007). Scale 1 is at the bottom, scale 3 at the top. Same color and number indicate same region in each scale.

However, standard CRF works on a very local level and long range dependencies are not addressed explicitly in simple CRF models. Therefore, our approach tries to set up a joint and hierarchical model of local and global information which explicitly models region adjacency graph (RAG) and region hierarchy graph (RHG) which is derived from a multi-scale image segmentation.

4.1 Proposed Model

Standard CRF acts on a local level and represents a single view on the data typically represented with unary and pairwise potentials. In order to overcome those local restrictions, we analyze the image at multiple scales $s \in \{1, ..., S\}$ with associated scale-specific unary potentials f_i^s and pairwise potentials f_{ij}^s, to enhance the model by evidence aggregation on local to global level. Furthermore, we integrate pairwise potentials g_{ik}^s to regard the hierarchical structure of the regions, i.e. if $i \in R^s$ then $k \in R^{s+1}$. In Fig. 1, we present a segmented image at three scales and the corresponding connectivity between the regions of successive scales. We see that regions that are too small to be classified accurately can inherit the labels of their parents. E. g. region 11 and 12 may be too small to reliably classify in isolation, but when they inherit a message from their parent region 5, they may possibly be correctly classified as 'cow'.

The proposed method explicitly models region adjacent neighborhood information within each scale or layer with f_{ij}^s and region hierarchical information between the scales with g_{ik}^s, using global image features as well as local ones for observations in the model. It has a distribution of the form

$$P(\mathbf{L} \mid \mathbf{X}) = \frac{1}{Z} \exp \left(\sum_{s=1}^{S} \sum_{i \in R^s} f_i^s(l_i \mid \mathbf{X}) \right.$$
$$\left. + \sum_{s=1}^{S} \sum_{(i,j) \in N^s} f_{ij}^s(l_i, l_j \mid \mathbf{X}) + \sum_{s=1}^{S-1} \sum_{(i,k) \in H^s} g_{ik}^s(l_i, l_k \mid \mathbf{X}) \right) \tag{2}$$

where R^s is the indexing set for regions corresponding to scale s, N^s is the set of neighboring regions at scale s, and H^s is the set of parent child relations between regions in neighboring scales s and $s+1$. Note that we use the same Z as the partition function for normalization as in standard CRF, although the value is different. We denote this model as Hierarchical Conditional Random Field (HCRF).

The proposed full graphical model is illustrated in Fig. 2. Note that this model only exploits up to second-order cliques, which makes learning and inference much easier. This model combines different views on the data by scale-specific potentials and the hierarchical structure accounting for longer range dependencies.

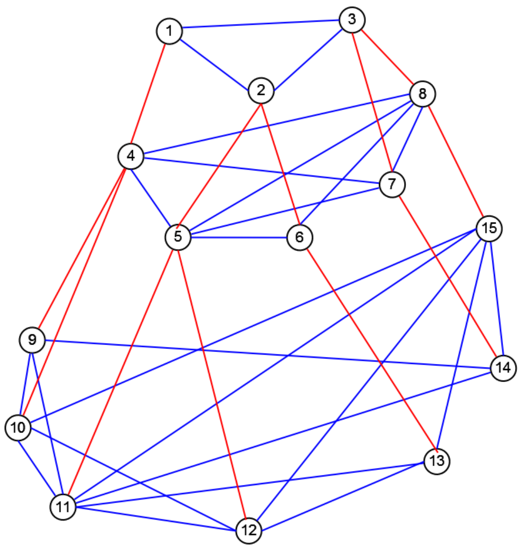

Figure 2: Illustration of the HCRF model architecture. The number of the nodes correspond to the regions in Fig. 1. The blue edges between the nodes represent the neighborhoods at one scale, the red edges represent the hierarchical relation between regions.

4.1.1 Unary Potentials

The local unary potentials f_i^s independently predict the label l_i based on the image $\mathbf{X}$:

$$f_i^s(l_i \mid \mathbf{X}) = \log P^s(l_i \mid \mathbf{X}). \tag{3}$$

The label distribution $P^s(l_i \mid \mathbf{X})$ is calculated by using a classifier. We employ the *multiple logistic regression model*,

$$P^s(l_i = c \mid u_{ic}^s) = \exp(u_{ic}^s) / \sum_{c'} \exp(u_{ic'}^s), \tag{4}$$

where $u_{ic}^s = \mathbf{w}_c^{s\mathsf{T}} \mathbf{h}_i^s$, $\mathbf{w}_c^s = [w_0^s, w_1^s, \ldots, w_M^s]$ are $M+1$ unknown parameters per class, and the feature vector $\mathbf{h}_i^s = [1, h_{i1}^s, \ldots, h_{im}^s, \ldots, h_{iM}^s]^\mathsf{T}$ contains M features for each region i derived from the image $\mathbf{X}$. The weights $\mathbf{w}^s = \{\mathbf{w}_c^s\}_{c=1,\ldots,C}$ are the model parameters.

4.1.2 Pairwise Potentials

The local pairwise potentials f_{ij}^s describe category compatibility between neighboring labels l_i and l_j given the image $\mathbf{X}$, which take the form of a contrast sensitive Potts model:

$$f_{ij}^s(l_i, l_j \mid \mathbf{X}) = \mathbf{v}^{s\mathsf{T}} \boldsymbol{\mu}_{ij}^s \delta(l_i \neq l_j) \tag{5}$$

where the feature function $\boldsymbol{\mu}_{ij}^s$ relate to the pair of regions (i, j), and the weights $\mathbf{v}^s$ again are the model parameters.

The hierarchical pairwise potentials g_{ik}^s also describe category compatibility between hierarchically neighboring labels l_i and l_k given the image $\mathbf{X}$, which take the form of a contrast sensitive Potts model:

$$g_{ik}^s(l_i, l_k \mid \mathbf{X}) = \mathbf{r}^{s\mathsf{T}} \boldsymbol{\eta}_{ik}^s \delta(l_i \neq l_k) \tag{6}$$

where the feature function $\boldsymbol{\eta}_{ik}^s$ relate to the hierarchical pairs of regions (i,k), and the vector $\mathbf{r}^s$ contains the model parameters. We denote the unknown HCRF model parameters by $\theta = \{\mathbf{w}^s, \mathbf{v}^s, \mathbf{r}^s\}_{s=1,\ldots,S}$.

4.2 Generating Multi-scale Segmentations

We now explain how we realized the multi-scale image segmentation and how we generate the region adjacency graphs (RAG) and region hierarchy graph (RHG).

We determine the image segmentation from the watershed boundaries on the image's gradient magnitude. Our approach uses the Gaussian scale-space for obtaining regions at several scales. The segmentation procedure has been described in detail by Drauschke et al. (2006). For each scale s, we convolve each image channel with a Gaussian filter and combine the channels when computing the gradient magnitude. Since the watershed algorithm is inclined to produce over-segmentation, we suppress many gradient minima by resetting the gradient value at positions where the gradient is below the median of the gradient magnitude. So, those minima are removed, which are mostly caused by noise. As a result of the watershed algorithm, we obtain a complete partitioning of the image for each scale s, where every image pixel belongs to exactly one region. Additionally, we determine the scale-specific RAGs on each image partition.

The development of the regions over several scales is used to model the RHG. Drauschke (2009) defined a RHG with directed edges between regions of successive scales (starting at the lower scale). Furthermore, the relation is defined over the maximal overlap of the regions. This definition of the region hierarchy leads to a simple RHG. If the edges would be undirected, the RHG only consists of trees.

4.3 Parameter Learning and Inference

For parameter estimation we take the learning approach (Sutton and McCallum, 2005) assuming the parameters of unary potentials to be conditionally independent of the pairwise potentials' parameters, allowing separate learning of the unary and the binary parameters. Note this no longer guarantees to find the optimal parameter setting for θ. In fact, the parameters are optimized to maximize a lower bound of the full CRF likelihood function by splitting the model into disjoint node pairs and integrating statistics over all of these pairs. Prior to learning the pairwise potential models we train parameters $\{\mathbf{w}^s\}_{s=1,\dots,S}$ for the unary potentials. Then, the pairwise potentials' parameter sets $\{\mathbf{v}^s\}_{s=1,\dots,S}$ and $\{\mathbf{r}^s\}_{s=1,\dots,S}$ are learned jointly in a maximum likelihood setting with stochastic meta descent Vishwanathan et al. (2006). We also assume a Gaussian prior on the linear weights to avoid overfitting (Vishwanathan et al., 2006).

We use max-product propagation inference (Pearl, 1988) to estimate the max-marginal over the labels for each region, and assign each region the label which maximizes the joint assignment to the image.

4.4 Feature Functions

To complete the details of our method, we now describe how the feature functions are constructed from low-level descriptors. They link the potentials to the actual image evidence and account for local neighborhood and long range dependencies.

Unary feature function $\mathbf{h}_i^s$ is a function of a predefined description vector for each region i at scale s.

Local pairwise potentials are responsible for modeling local dependencies by supporting or inhibiting label propagation to the neighboring regions. Therefore, we define the local pairwise function $\boldsymbol{\mu}_{ij}^s$ as

$$\boldsymbol{\mu}_{ij}^s = \left[1, \{|h_{im}^s - h_{jm}^s|\}\right]^\top \tag{7}$$

Here, we extended each difference by an offset for being capable eliminating small isolated regions.

Hierarchical pairwise potentials act as a link across scale, facilitating propagation of information in our model. Therefore, we define the hierarchical pairwise function $\boldsymbol{\eta}_{ik}^s$ as

$$\boldsymbol{\eta}_{ik}^s = \left[1, \{|h_{im}^s - h_{km}^{s+1}|\}\right]^\top \tag{8}$$

where region i is at scale s and region k is at scale $s+1$.

In the following, we give an example of how we build the description vector for each region mentioned above in the context of building facade interpretation.

For each region i at the highest resolution, say, at scale with index 1, we compute an 75-dimensional description vector $\boldsymbol{\phi}_i^1$ incorporating region area and perimeter, its compactness and its aspect ratio. For representing spectral information of the region, we use same 12 color features as Barnard et al. (2003): the mean and the standard deviation of the RGB and the Lab color spaces. We also include features derived from the gradient histograms as it has been proposed by Korč and Förstner (2008). Additionally we use texture features derived from the Walsh transform (Petrou and Bosdogianni, 1999; Lazaridis and Petrou, 2006). Other features are derived from generalization of the region's border and represent parallelity or orthogonality of the border segments, or they are descriptors of the Fourier transform.

We define this description vector to be the unary feature function $\mathbf{h}_i^1$ at scale 1. For the higher scales s, we compute the description vector $\boldsymbol{\phi}_i^s$ and unary feature function $\mathbf{h}_i^s$ using the correspondent regions at lower scales.

We have finished the multi-scale image segmentation and feature extraction on eTRIMS database [1]. Based on segmented regions, we have generated RAG and RHG. We are currently working on learning and inference issues.

5 SUMMARY

In this paper, we have shown a novel approach called hierarchical conditional random field (HCRF). The proposed method explicitly models region adjacent neighborhood information within each scale and region hierarchical information between the scales, using global image features as well as local ones for observations in the model. This model only exploits up to second-order cliques, which makes learning and

[1] http://www.ipb.uni-bonn.de/projects/etrims/

inference much easier. This model combines different views on the data by layer-specific potentials and the hierarchical structure accounting for longer range dependencies.

REFERENCES

Barnard, K., Duygulu, P., Freitas, N. D., Forsyth, D., Blei, D., and Jordan, M. (2003). Matching Words and Pictures. In *JMLR*, volume 3, pages 1107–1135.

Drauschke, M. (2009). An Irregular Pyramid for Multi-scale Analysis of Objects and their Parts. In *7th IAPR-TC-15 Workshop on Graph-based Representations in Pattern Recognition*, pages 293–303.

Drauschke, M., Schuster, H.-F., and Förstner, W. (2006). Detectability of Buildings in Aerial Images over Scale Space. In *PCV'06*, IAPRS 36 (3), pages 7–12.

Gould, S., Rodgers, J., Cohen, D., Elidan, G., and Koller, D. (2008). Multi-Class Segmentation with Relative Location Prior. *IJCV*, 80(3):300–316.

He, X., Zemel, R., and Carreira-Perpin, M. (2004). Multi-scale Conditional Random Fields for Image Labeling. In *CVPR*, pages 695–702.

Kohli, P., Kumar, M. P., and Torr, P. (2007). P3 & Beyond: Solving Energies with Higher Order Cliques. In *CVPR*, pages 1–8.

Kohli, P., Ladicky, L., and Torr, P. (2009). Robust Higher Order Potentials for Enforcing Label Consistency. *IJCV*, 82(3):302–324.

Korč, F. and Förstner, W. (2008). Interpreting Terrestrial Images of Urban Scenes using Discriminative Random Fields. In *21st ISPRS Congress*, IAPRS 37 (B3a), pages 291–296.

Kumar, S. and Hebert, M. (2003). Discriminative Random Fields: A Discriminative Framework for Contextual Interaction in Classification. In *ICCV*, pages 1150–1157.

Ladicky, L., Russell, C., and Kohli, P. (2009). Associative Hierarchical CRFs for Object Class Image Segmentation. In *ICCV*, pages 1–8.

Lafferty, J., McCallum, A., and Pereira, F. (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *ICML*, pages 282–289.

Lazaridis, G. and Petrou, M. (2006). Image Registration using the Walsh Transform. *Image Processing*, 15(8):2343–2357.

McCallum, A., Rohanimanesh, K., and Sutton, C. (2003). Dynamic Conditional Random Fields for Jointly Labeling Multiple Sequences. In *NIPS Workshop on Syntax, Semantics and Statistic*.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann.

Petrou, M. and Bosdogianni, P. (1999). *Image Processing: The Fundamentals*. Wiley.

Plath, N., Toussaint, M., and Nakajima, S. (2009). Multi-Class Image Segmentation using Conditional Random Fields and Global Classification. In *ICML*, pages 817–824.

Reynolds, J. and Murphy, K. (2007). Figure-ground segmentation using a hierarchical conditional random field. In *4th Canadian Conference on Computer and Robot Vision*, pages 175–182.

Schnitzspan, P., Fritz, M., Roth, S., and Schiele, B. (2009). Discriminative Structure Learning of Hierarchical Representations for Object Detection. In *CVPR*, pages 2238–2245.

Schnitzspan, P., Fritz, M., and Schiele, B. (2008). Hierarchical Support Vector Random Fields: Joint Training to Combine Local and Global Features. In *ECCV*, pages 527–540.

Shotton, J., Winnand, J., Rother, C., and Criminisi, A. (2006). Textonboost: Joint Appearance, Shape and Context Modeling for Multi-Class Object Recognition and Segmentation. In *ECCV*, pages 1–15.

Sutton, C. and McCallum, A. (2005). Piecewise Training for Undirected Models. In *21th Ann. Conf. on Uncertainty in AI*, pages 568–575.

Toyoda, T. and Hasegawa, O. (2008). Random Field Model for Integration of Local Information and Global Information. *PAMI*, 30(8):1483–1489.

Vishwanathan, S. V. N., Schraudolph, N. N., Schmidt, M. W., and Murphy, K. P. (2006). Accelerated Training of Conditional Random Fields with Stochastic Gradient Methods. In *ICML*, pages 969–976.

Yang, L., Meer, P., and Foran, D. J. (2007). Multiple Class Segmentation using a Unified Framework over Mean-Shift Patches. In *CVPR*, pages 1–8.