Multi-View Random Fields and Street-Side Imagery

Michal Recky
ICG TUGraz
Graz University of Technology
Inffeldgasse 16
A-8010 Graz, Austria
recky@icg.tugraz.at

Franz Leberl
ICG TUGraz
Graz University of Technology
Inffeldgasse 16
A-8010 Graz, Austria
leberl@icg.tugraz.at

Andrej Ferko
FMFI UK
Mlynská dolina
842 48 Bratislava, Slovakia
ferko@sccg.sk

ABSTRACT

In this paper, we present a method that introduces graphical models into a multi-view scenario. We focus on a popular Random Fields concept that many researchers use to describe context in a single image and introduce a new model that can transfer context directly between matched images – Multi-View Random Fields. This method allows sharing not only visual information between images, but also contextual information for the purpose of object recognition and classification. We describe the mathematical model for this method as well as present the application for a domain of street-side image datasets. In this application, the detection of façade elements has improved by up to 20% using Multi-view Random Fields.

Keywords

Random Fields, Context, Redundancy, Image Stacks

1. INTRODUCTION

In a current computer vision research input data is often represented as large, redundant datasets with hundreds or even thousands of overlapping images. As the volume and complexity of data increases, it is no longer meaningful to employ manual inputs in any step of the process. This constraint on the work automation leads to a need to utilize as much information from images as possible. One potential approach is to employ "context". Most popular methods of context application are graphical models, specifically Random Fields. However, general Random Fields models are defined such that they allow observations only from a single image. This approach is limiting context as a feature of a single image, but the context is derived from objects in a real scene, from which an image is only one projection. How is this limiting context application and how can we expand the Random Fields model to cope with the presence of multi-view dataset is the topic of this paper.

The basic element in a Random Field model is a "site". This is generally a patch of image area ranging from a single pixel to a larger segment. In our application in a street-side images domain, a site is a rectangular area (block) of a building façade (see Figure 1). Each site has to be labeled according to visual data and a context in which it is observed. Context is defined as relations (spatial relations, similarity…) between sites. In a multi-view scenario, we have multiple matched images, each with its own set of sites. Extension of Random Fields into a multi-view is not straightforward, as the two sets of sites from matched images are typically overlapping.

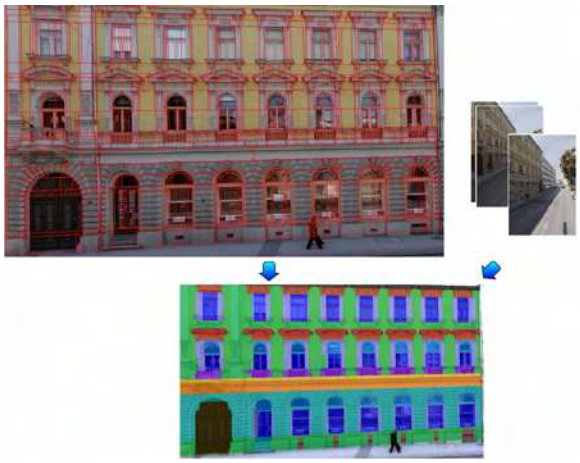

Figure 1: The application of Multi-View Random Fields for labeling of the façade elements. Top left – set of blocks that divide building façade into a set of sites for a graphical model. Bottom – final labeling is achieved as a combination of information from multiple overlapping images (for color-coding, see Figure 7).

Simple merging of these two sets would cause double detections of same objects and unresolved relations between sites. To solve both problems, we introduce a new concept – Multi-View Random Fields.

In this paper, the "*Background"* and "*Graphical Models"* sections are outlining a context of our work in a computer vision community and in a Random Fields models research. The "*Context in Multi-View"* section explains what type of context is relevant in multi-view and how it can be utilized. In the "*Multi-View Random Fields"* section the new graphical model is introduced and the *"Application of MVRF"* section present the illustrational application of the model in a street-side images domain.

2. BACKGROUND

The last decade saw growing interest in multi-view methods. With the introduction of a new generation of high resolution digital cameras and with rapid improvements in storage and computing hardware, multi-view imagery advanced from a source for the computation of point clouds by two-image stereo methods to a broad range of vision problems employing thousands of overlapping images. Open online image hosting sites (Flickr, Picasa, Photobucket…) have added interesting vision opportunities. While the basic principles for matching images remain applicable to such datasets [Har04a] [Leo00a], new problems needed to get solved, such as the organization and alignment of images without any knowledge about camera poses [Sna06a]. The resulting resource need in computing gets addressed by means of graphical processing units GPUs, or with distributed approaches [Fra10a]. Therefore current computer vision can cope with this avalanche of imagery and multi-views are becoming a common reality.

Extending the concept of Random Fields into such multi-view scenario comes from an idea that given more images of the same scene, more contextual relations can be examined. In this work, we present a mathematical model for Multi-View Random Fields that allows transferring contextual relations between matched images. We also present the application of Multi-View Random Fields in a domain of street-side images. This domain is useful for a demonstration, as there are large datasets of matched street-side images for the purpose of urban modeling (virtual cities, GIS, cultural heritage reconstruction) that establish a multi-view scenario. Urban scenes also exhibit strong contextual relations, as man-made objects adhere to an inherent organization. We show how façade elements can be classified, using both context and multi-view principles in one model.

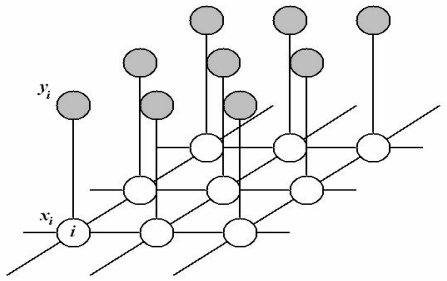

Figure 2. The typical application of MRF in computer vision. At each node (site) *i*, the observed data is denoted as y_i and the corresponding label as x_i. For each node, only local observations are possible. Generally each node represents a pixel in an image and observed data pixel's features.

3. GRAPHICAL MODELS

The most common non-causal graphical models in computer vision are Markov Random Fields (MRF).

MRF have been used extensively in labeling problems for classification tasks in computer vision [Vac11a] and for image synthesis problems. In a labeling task, MRF are considered to be probabilistic functions of observed data in measured sites of the image and labels assigned to each site. Given the observed data $\mathbf{y} = \{\mathbf{y}_i\}_{i \in S}$ from the image, and corresponding labels $\mathbf{x} = \{x_i\}_{i \in S}$, where S is the set of sites, the posterior distribution over labels for MRF can be written as:

$$P(\mathbf{x} \mid \mathbf{y}) = \frac{1}{Z_m} \exp\left(\sum_{i \in S} \log p(\mathbf{y}_i \mid x_i) + \sum_{i \in S} \sum_{j \in N_i} \beta_m x_i x_j \right), \quad (1)$$

where Z_m is the normalizing constant, β_m is the interaction parameter of the MRF and N_i is the set of neighbors of a site i. The pairwise term $\beta_m x_i x_j$ in MRF can be seen as a smoothing factor. Notice that the pairwise term in MRF uses only labels as variables, but not the observed data from an image. In this arrangement, the context in a form of MRF is limited to be a function of labels, thus allowing for semantic context (context between classes) and limiting geometric context to a structure of MRF graph (see Figure 2). This makes the MRF applicable mainly for simpler forms of local context.

To cope with such limitations, the concept of Conditional Random Fields (CRF) was proposed by J. Lafferty [Laf01a] for the segmentation and labeling of text sequences. The CRF are discriminative models that represent the conditional distribution over labels. Using the Hammersley-Clifford theorem [Ham71a], assuming only pairwise cliques potentials to be nonzero, the conditional distribution in CRF over all labels **x** given the observation **y** can be written as

$$P(\mathbf{x} \mid \mathbf{y}) = \frac{1}{Z} \exp\left(\sum_{i \in S} A_i(x_i, \mathbf{y}) + \sum_{i \in s} \sum_{j \in N_i} I_{ij}(x_i, x_j, \mathbf{y}) \right), \quad (2)$$

where Z is the normalizing constant, $-A_i$ is the unary and $-I_{ij}$ pairwise potential. The two principal differences between conditional model (2) and MRF distribution (1) are that the unary potential $A_i(x_i, \mathbf{y})$ is a function of all observations instead of only one observation $\mathbf{y}_i$ in a specific site i and the pairwise potential in (2) is also the function of observation, not only labels as in MRF. In CRF, the unary potential $A_i(x_i, \mathbf{y})$ is considered to be a measure of how likely a site i will take label x_i given the observation in a image $\mathbf{y}$. The pairwise term is considered to be a measure of how the labels at neighboring sites i and j should interact given the observed image $\mathbf{y}$. This concept of CRF allows for use of more complex context derived from larger sets of observations in the image and employing geometric context (e.g. spatial relations between objects). It is extended even more in a concept of Discriminative Random Fields [Kum06a], where an arbitrary discriminative classifier can be applied in a form of unary/pairwise potential.

However, in all concepts of Random Fields, the set of sites S (and thus the observations) is limited to a single image. How to extend these models into a multi-view is explained in subsequent sections.

4. CONTEXT IN MULTI-VIEW

Before the definition of a new Random Field model in multi-view, we must consider what type of context can be transferred between images. The most common type of context applied for classification is a local pixel context. In general, a small neighborhood around an examined pixel is taken as a context area and a graph structure of a model is placed in this neighborhood (one node per pixel). However, this approach is not suitable for multi-views, as neighborhoods around matched pixels in two images are in general uniform and will not present much useful additional information. Alternatively we can consider global context, which examines relationships between all objects in the images. In this type of context, we can observe different relations in different images, thus transferring such context would provide additional information for recognition and classification (see Figure 3). If spatial relations between objects are examined in this manner, graphical models are approximating spatial relations between objects in a real 3D scene.

In a standard Random Fields (RF) model, each image is considered a unique unit of information. Thus, we can consider a global context to be a specific feature of each image - the global context is a set of relations between all sites detected in a single image.

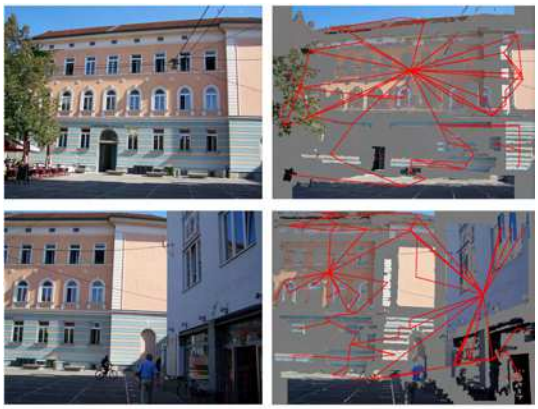

Figure 3. Building façade projected in slightly different views. Red lines (graph edges) represent spatial relationships between objects detected in the images, indicating different context in two projections for the same objects. For better overview, only some relations are visualized.

Typically, sites are either *pixels* or *segments*. Construction of a global model with node in each pixel would significantly increase the complexity of computation; therefore we consider segments as the representation of sites in our model.

Subsequently a site is represented by a specific area (segment) in a digital image. Such area represents an object (or part of object) and areas from two sites are not overlapping. In a general RF model, a set of all sites in one graph is denoted as S. In a local model, one set S include sites from a small patch of the image, however in a global model, S includes all sites from the entire image. Visual features of the area assigned to a specific site are denoted as image observation $\boldsymbol{y}_s$ from site $s \in S$. In a graphical model, if there is an edge between nodes assigned to sites s_1 and s_2, let's denote this relation as $\Phi(s_1, s_2) = 1$ and consequently if there is no edge between s_1 and s_2, denote this as $\Phi(s_1, s_2) = 0$.

Transferable Sites

Consider one image from the dataset as "examined image" to which we would like to transfer context from other matched images. Let's call any site $s \in S$ from an examined image a "native site". If the image matching is established in a dataset (we have a set of corresponding points that link images), we can look for any sites from other images that are *corresponding* to native sites. In most cases, sparse point cloud of matched points is enough to establish correspondence between site. Relative poses between images and camera parameters are not required. Definition of corresponding sites can vary in different applications. In general, *corresponding sites are two sites from different images that share some subset of corresponding points*;

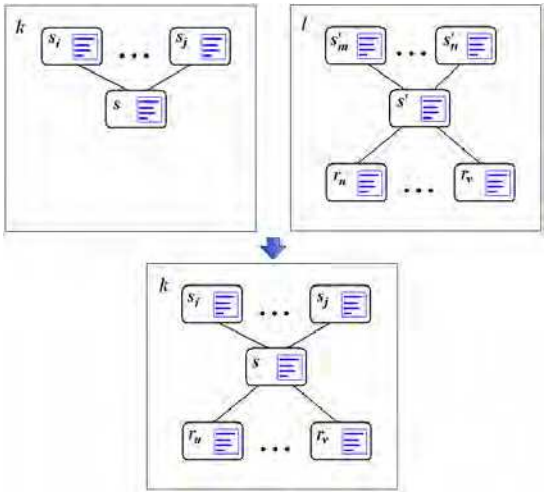

Figure 4. Transfer of sites from the image $l \in$ I to the image $k \in$I, as presented in Definition 1. Only sites from *l* that are not corresponding to any sites from *k* are transferred. This figure demonstrates only transfer between two images.

each site from matched images can have only one corresponding site in the examined image – the example of this relation is provided in the application section of this paper.

Given that corresponding sites usually represent the same objects, transferring such information between images would be redundant. Therefore we transfer sites that have no correspondences in the examined images to provide new information. We denote such sites as "transferable sites". For a single, examined image from the image stack, let's define the set of transferable sites as:

Definition 1: *If* $S_k = \{s_1, s_2, \ldots, s_n\}$ *is the set of sites for single image* $k \in$ I, *where* I *is the set of images and correspondences have been established between the images from* I *such that* $s_i' \in S_l$ *is a site from image* $l \in I-\{k\}$ *corresponding to a site* s_i. *Than the* $R_k = \{r_1, r_2, \ldots, r_m\}$ *is the set of transferable sites for the image k if* $\forall r_j \in R_k \exists s_i \in S_k \mid \Phi(r_j, s_i') = 1$ *and* $\forall r_j \in R_k \neg \exists r_j' \in S_k$. R_k *is constructed such that* $\forall r_i, r_j \in R_k$, r_i *and* r_j *are not correspondent to each other in any two images from I*

Thus the R_k is the set of sites from other images than *k*, that are in the relationship in graphical model with some corresponding site to sites from S_k, but themselves have no correspondences in S_k (see Figure 4). The set of transferable sites can be seen as a context information, that is available in the image stack, but not in the examined image. If sites are the representations of objects, than in a transferable set, there are objects in context with the scene of the image that are currently not located in the projection, thus are occluded, out of the view or in different timeframe. This also means that the visual information from the sites in R_k are not present in the image *k*. If the sites from R_k are included in the vision process, they can provide additional context and visual information that is not originally present in the examined image.

Note that a transferable site is not equivalent to a native site in an examined image. Even though transferable sites have the same set of visual features as sites native to the image and they can be assigned the same set of spatial and contextual relations in a graphical model, transferable sites lost all original contextual relationships except the relationships to the sites they are connected within the examined image. This makes them harder to label. But the labeling of transferable sites is not the aim in the case of examined image (the goal is to label only native sites), thus transferable sites can contribute information for image labeling, but the labeling of themselves is usually irrelevant.

5. MULTI-VIEW RANDOM FIELDS

Given a non-equality of transferable sites to native sites, standard RF models are not compatible with this extended set. For this reason, we introduce a new model denoted as *Multi-View Random Fields* (MVRF). This model is derived from a CRF, described in Section 2; however we extend the posterior probability distribution into MVRF model framework as follows:

Given the observed data $\mathbf{y} = \{\mathbf{y}_i\}_{i \in S}$ from the image, corresponding labels $\mathbf{x} = \{x_i\}_{i \in S}$, where S is the set of native sites from the image and observations from transferable set $\mathbf{z} = \{\mathbf{z}_i\}_{i \in R}$ with corresponding labels $\tilde{\mathbf{x}} = \{\tilde{x}_i\}_{i \in R}$, where R is the set of transferable sites, the posterior distribution over labels is defined as:

$$P(\mathbf{x} \mid \mathbf{y}, \mathbf{z}) = \frac{1}{Z} \exp\left(\sum_{i \in S} A_i(x_i, \mathbf{y}) + \sum_{i \in R} A_i'(\tilde{x}_i, \mathbf{z}_i) + \right.$$
$$\left. + \sum_{i \in S} \left(\sum_{j \in N_i} I_{ij}(x_i, x_j, \mathbf{y}) + \sum_{j \in K_i} I_{ij}'(x_i, \tilde{x}_j, \mathbf{y}, \mathbf{z}_j) \right) \right) \quad ,(3)$$

where *Z* is the normalizing constant, N_i is the set of native sites neighboring site *i* and K_i is the set of transferable sites neighboring site *i*. - A_i and - A_i' are unary potentials, - I_{ij} and - I_{ij}' are pairwise potentials (for native sites and transferable sites respectively). The differences between potentials for transferable sites and for native sites are as follows:

- In the unary potential for a transferable site, only observations from the site itself are

considered, instead of observation from the entire image for native sites. This is due to the fact, that a transferable site does not have any connections to the image except for the site it is neighboring. Even if other connections exist (with other sites in the image), it is a hard task to establish relationships. For native site, there are no changes to a standard conditional model.

- In the pairwise potential, in addition to observation from the image, local observation from the transferable site is considered, when relations are examined between a native site and transferable site. The inclusion of all image observation grant at least the same level of information in pairwise computation as in a standard CRF model and the additional observation form transferable site represent extended context for native image observation. The pairwise potential for two native sites is the same as in a standard CRF model.

This model has some additional unique characteristics. For example, no pairwise relations are considered between two transferable sites. This is based on the construction of transferable sites set. A site from such set can be neighboring several native sites, but not any other transferable site. This can be seen as a limitation for the model, however without additional high frequency information about the scene (as a prior knowledge), it is virtually impossible to establish relationships for transferable sites.

The computational complexity of the model is not increased significantly. Pairwise potentials are computed only for native sites, as it is in the standard CRF model. The difference is in the number of neighbors for each site, however even this number should not increase significantly. When considering a global model, each new neighbor (transferable site in relation to the native site) represents a new object in the projection. This is dependent on the differences between projection parameters – camera positions, optical axes…, but even for very different parameters, the number of objects should not differ significantly for the same scene. From the general observation, the number of neighboring transferable sites is notably lower than the number of neighboring native sites.

Potentials Modifications

Unary potential for native image sites, similar to a standard CRF is a measure of how likely a site i will take label x_i given the observations in image $\mathbf{y}$. A standard approach described in a work of S. Kumar is to apply Generalized Linear Models (GLM) as local class conditional [Kum06]. In that case, given the model parameter $\mathbf{w}$ and a transformed feature vector at each site $\boldsymbol{h}_i(\mathbf{y})$, the unary potential can be written as:

$$A_i(x_i, \mathbf{y}) = \boldsymbol{log}\left(\sigma\left(x_i \mathbf{w}^T \boldsymbol{h}_i(\mathbf{y})\right)\right) \quad ,(4)$$

For the transferable sites, the feature vector is limited to the observations from single site. This limitation defines a new expression for unary potential, exclusive to transferable sites as

$$A_i'(\tilde{x}_i, \mathbf{z}_i) = \boldsymbol{log}\left(\sigma\left(\tilde{x}_i \mathbf{w}^T \boldsymbol{h}_i(\mathbf{z}_i)\right)\right) \quad ,(5)$$

The feature vector $\boldsymbol{h}_i(\mathbf{z}_i)$ at the transferable site i is defined as a nonlinear mapping of site feature vectors into high dimensional space. The model parameter $\mathbf{w} = \{w_0, \mathbf{w}_1\}$ is composed of bias parameter w_0 and model vector $\mathbf{w}_1$. $\sigma(.)$ is a local class conditional, that can be any probabilistic discriminative classifier.

The pairwise potential for two native sites from the image remains the same as in CRF model, given the GLM are applied to compute the class conditional:

$$I_{ij}(x_i, x_j, \mathbf{y}) = \beta\left(K x_i x_j + (1-K)\left(2\sigma\left(x_i x_j \mathbf{v}^T \boldsymbol{\mu}_{ij}(\mathbf{y})\right) - 1\right)\right) \quad ,(6)$$

where $0 \le K \le 1$, $\mathbf{v}$ and β are the model parameters and $\boldsymbol{\mu}_{ij}(\mathbf{y})$ is a feature vector. For transferable sites, we introduce the additional feature vector in a form of observations from specific site:

$$I_{ij}'(x_i, \tilde{x}_j, \mathbf{y}, \mathbf{z}_j) = \beta\left(K x_i \tilde{x}_j + (1-K)\left(2\sigma\left(x_i \tilde{x}_j \mathbf{v}^T \boldsymbol{\mu}_{ij}(\mathbf{y}, \mathbf{z}_j)\right) - 1\right)\right) \quad ,(7)$$

where $\mu_{ij}(\mathbf{y},\mathbf{z}_i)$ is a feature vector defined in a domain $\mu : \Re^{\gamma} \times \Re^{\gamma} \to \Re^{q}$ such that observations are mapped from the image/sites related to site s into a feature vector with dimension γ. Note that the smoothing term $K x_i \tilde{x}_j$ is the same as in a standard CRF definition. Thus if $K = 1$, the pairwise potential still performs the same function as in a MRF model, however given new transferable sites, the smoothing function will depend also on their classification $\tilde{x}_j$.

In this case, visual information from transferable sites is not involved in the pairwise term and is only applied in the unary term. If $K<1$ the data-dependent term $2\sigma\left(x_i \tilde{x}_j \mathbf{v}^T \boldsymbol{\mu}_{ij}(\mathbf{y}, \mathbf{z}_j)\right) - 1$ is included in a pairwise potential. Observations from the image related to the examined native site and observation from transferable site are transformed into feature vector and involved in computation.

Parameter Learning and Inference

In this work, we constructed an MVRF model to be as compatible with other RF models as possible. This approach is observed also in a parameter learning process, as any standard method used for learning of

CRF model can be also used for MRVF model. To further simplify the process, we observed that learning from single (un-matched) images is feasible without the loss of strength of the model. This is due to the construction of potentials - in a unary potential, visual features do not change for transferable sites, therefore they can be learned directly from single images in training dataset. The spatial relations defined for a pairwise potential also do not change significantly for the pair native-transferable site. For such reasons, we can assume that the MVRF model can be learned even directly from single images without dataset matching. Therefore, methods such as pseudo-likelihood can be applied for learning.

Similarly, parameter inference can be performed, using any standard method applied in CRF. In our application, we use Belief Propagation, but other possible methods are Tree-Based Reparameterization or Expectation Propagation for example.

6. APPLICATION OF MVRF

In this section we present the application of MVRF in the building façades dataset for the purpose of façade elements detection and classification. This application is based on the dataset provided by a vehicle-based urban mapping platform. Sparse image matching is applied (see Figure 5), using the Structure-from-Motion method [Irs07a]. We selected the left camera subset, since it provides a clear view of the building façades, not distorted by the perspective (which, however, is easy to rectify) and with good visual cues. This setting will demonstrate the advantages of MVRF in cases when a site was misdetected and presents lost contextual information in standard models. In most images, the building façade is not projected in its entirety and parts are located in other images. Therefore in such cases, the MVRF will also provide new contextual and visual information in a form of transferable sites based on the objects that are not located in the original image.

In each image, separate facades are detected. This can be achieved when the wire-frame models of the scene are available, or using visual cues, such as repetitive patterns [Rec11a]. Subsequently, a modified gradient projection is applied to segment each façade into a set of blocks. This method is based on a standard gradient projection approach [Lee04a] designed for the detection of windows with following modifications:

First, we vertically project gradients to establish a horizontal division of the façade into levels (level borders are located at the spikes of the projection). Subsequently, we compute horizontal gradient projections in each level separately.

Figure 5. Top row: two examples of the same façade, matched with a sparse point cloud (red dots). Middle row: set of blocks located in each façade (left image show façade detail for better overview, right image entire facade). Bottom row: set of blocks from the first image projected into a second image and a set of transferable sites (highlighted blocks) that is derived from the projection (as sites that have no correspondence in second set).

This process will yield a set of blocks bordered by level borders horizontally and spikes in projection vertically (see Figure 5). Second, we consider each block as a site for a graphical model, thus we compute visual features for each block and consider spatial relationships between blocks. Visual features, such as texture covariance, or clustering in a color space are used for classification [Rec10a]. For example, clusters in a CIE-Lab color space are computed for each block and are compared to class descriptors.

When the segmentation of a façade into a set of block is established, we can define a global graphical model in this structure. Each block is considered a site, thus each node of the graph is placed in a separate block. We define neighborhood relation such that for each block, its neighbors are all blocks located in areas above, below, left and right from itself (see Figure 6). This definition allows considering all objects at the same level and column to be involved in contextual relations, accounting for relations, such as rows and columns of windows, or window-arch. An edge of a graphical model is placed between each two neighboring blocks. In this approach, a separate graph is created for each façade in the image.

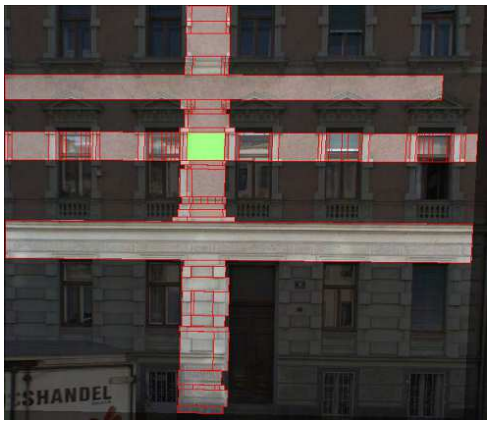

Figure 6. Example of site neighborhood, as defined in this application. Green block is the examined site and highlighted blocks are defined as its neighborhood.

Multi-View Scenario

To establish a multi-view, we use a sparse point cloud. We match blocks between images such that we interpolate between detected corresponding points to achieve rough point-to-point matching. If two blocks in different images share at least 2/3 of matched points (detected and interpolated), we define these as corresponding blocks. Given one image as "examined", we can label all blocks from the same façade in other images as either corresponding or non-corresponding. Subsequently, transferable sites are blocks that are from the same façade as in an examined image, but are non-corresponding to any block from the examined set (see Figure 5). Establishing the relations between native and transferable sites is straightforward, as we can still consider up, down, left, right directions. With these definitions, we can construct the MVRF model from our dataset.

Experiments

We use the described model for the purpose of façade elements detection and classification. The set of classes with corresponding color coding is displayed in Figure 7. Our testing dataset consists of 44 matched images. This dataset covers three full building façades and one half façade. A sparse point cloud of 1429 3D points is used to match images. Approximately 800 points are projected into each image. In the testing process, we compare the number of façade elements to the number of detected elements with the applied method. We counted overall numbers of elements through the entire dataset, as displayed in Table 1. For example, total number of 536 "window centre" elements can be observed in all images, that is approximately 12 "window centers" per image.

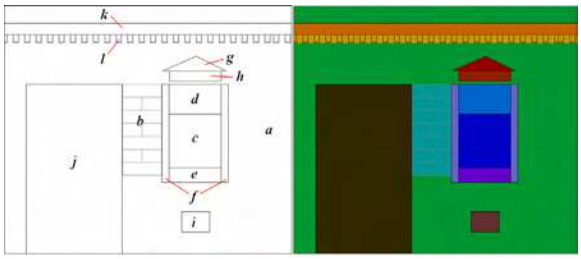

Figure 7. Set of classes: *a*) clear façade; *b*) brick façade; *c*) window centre; *d*) window top; *e*) window bottom; *f*) window margins; *g*) arch top; *h*) arch bottom; *i*) basement window; *j*) door; *k*) ledge; *l*) ledge ornament; On the right side, color representation of each class is displayed.

Each façade was processed separately, that is if there were two façades in one image, such image was processed two times (each time for different façade). After running the algorithm, a number of detected elements is counted visually. The façade element is defined as *detected*, if at least 2/3 of its area is labeled with the corresponding class. For the training purpose, we used the subset of 3 images from the dataset and other 5 unrelated images as labeled ground truth. This proved to be sufficient, as the spatial relations between classes are in general stable through different facades and a certain visual features variability

Class	# el	single	multi /native	multi /trans
clear façade	61	61	61	61
brick façade	54	54	54	54
win. centre	536	485	531	531
window top	311	270	303	308
win. bottom	300	227	273	288
win. margin	683	572	618	654
Arch top	199	176	189	192
Arch bottom	199	184	194	194
Basem. win	121	98	115	117
Door	34	32	33	33
Ledge	90	90	90	90
Ledge orna.	34	32	34	34

Table 1. The Results for the MVRF application. "# el" displays the overall number of each class for entire dataset (44 images). "single" displays detected elements in MVRF single image scenario (equivalent to CRF), "multi/native" displays results for multi-view scenario with only native sites in results and "multi/trans" display results for multi-view scenario with transferable sites labels in results. Numbers displayed are the detected façade elements in all images of dataset.

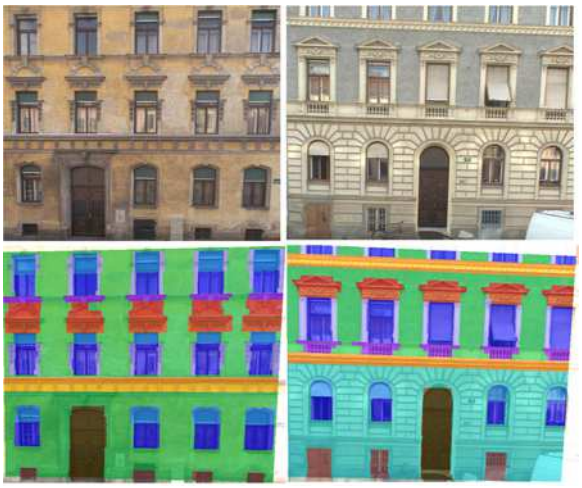

Figure 8. Two examples of classification results. Classes are labeled according to color scheme explained in Figure 7. Colors are superimposed over original images in the bottom row.

was allowed by the use of descriptors (e.g. clustering). We trained on single images without the use of matching. For the parameter inference, we used a Belief Propagation method. Initial classification was performed based on only visual features and in each iterative step of the method, it was refined by pairwise relations and site features described in a model. In each step, we also refined visual descriptors for each class to better approximate features in each unique façade. Results can be observed in the Table 1. We included results for scenarios, where no transferable sites were used (single), and the MVRF model is equivalent to CRF in this case, results when only labels of native sites were considered and results were labels of transferable sites were included. Notice a significant improvement in detection for classes that are visually ambiguous, but have strong contextual relations (e.g. window margins, window tops). For a "win. bottom" class, the correct detection rate improved from 76% in a single-view to a 96% in a multi-view with transferable sites projected, thus achieving a 20% improvement. Results illustrated in Figure 8.

7. CONCLUSION

In this paper, we addressed a common problem in a current research – how to work with context information in matched datasets and to alleviate an artificial limitation of graphical models to single images. We introduced a new MVRF model directly applicable in a multi-view scenario. We extended the standard CRF model such that it can work with overall context of the scene present in the multi-view dataset, but it still retains the same properties for processing visual and contextual information in a single image. Validity of this model is subsequently demonstrated in the application in street-side image domain – detection of façade elements. However the new MVRF model is applicable in same situations as a standard CRF model, provided that appropriate image matching is available. For example, the MVRF model was also used for a super-pixel based semantic segmentation of outdoor images in our other work.

8. REFERENCES

[Fra10a] Frahm, J. M., et al. Building Rome on a Cloudless Day. *European Conference on Computer Vision,* pp. 368–381, 2010

[Ham71a] Hammersley, J. M., Clifford, P. *Markov* field on finite graph and lattices. Unpublished, 1971

[Har04a] Hartley, R. and Zisserman, A. Multiple View Geometry in Computer Vision. *Cambridge University Press*, ISBN: 0521540518, 2004

[Irs07a] Irschara, A., et al. Towards wiki-based dense city modeling. *International Conference on Computer Vision*, pp. 1-8, 2007

[Kum06a] Kumar, S. and Herbert, M. Discriminative random fields. *International Journal of Computer Vision*, 68(2), pp. 179–201, 2006

[Laf01a] Lafferty, J., et al. Conditional Random Fields: Probabilistic models for segmenting and labeling sequence data. *International Conference on Machine Learning*, pp. 282-289, 2001

[Lee04a] Lee, S. C. and Nevatia, R. Extraction and integration of window in a 3d building model from ground view images. *Computer Vision and Pattern Recognition*, pp. 112-120, 2004

[Leo00a] Leonardis, A., et al. Confluence of computer vision and computer graphics. NATO science series, *Kluwer Academic Publishers*, ISBN 0-7923-6612-3, 2000

[Rec11a] Recky, M,. et al. Façade Segmentation in a Multi-View Scenario. *International Symposium on 3D Data Processing, Visualization and Transmission*, pp. 358-365, 2011

[Rec10a] Recky, M. and Leberl, F. Windows Detection Using K-means in CIE-Lab Color Space. *Internation Conference on Pattern Recognition,* pp. 356-360, 2010

[Sna06a] Snavely, N., et al. Photo tourism: Exploring photo collections in 3d. *ACM Transactions on Graphics*, pp. 835 – 846, 2006

[Vac11a] Vacha, P., et al. Colour and rotation invariant textural features based on Markov random fields. *Physical Review Letters, No. 6*, pp. 771-779, 2011