Preprint for CGIM 2010 Innsbruck

# MULTI-VIEW IMAGE INTERPRETATION FOR STREET-SIDE SCENES

Michal Recky, Franz Leberl
Institute for Computer Graphics and Vision
Graz University of Technology
Inffeldgasse 16
A-8010 Graz, Austria
recky@icg.tugraz.at

**ABSTRACT**
In this paper we examine the concept of redundancy and how it can improve the scene interpretation. In our work, we focus on redundant sets of street-side images. Semantic segmentation is performed on each image. Results of the segmentation are compared in overlapping images and matched. We use two principally different datasets to validate our results. The Industrial System dataset is taken from a moving car by well-designed, calibrated, automated cameras, with the geometry and pattern of the images accurately defined. Our second dataset (Tummelplatz-Graz) was taken by a hand-held camera in an urban environment, following the "crowd-sourcing" paradigm. Each database provides its typical level of redundancy and different approaches are needed for image matching. The annotated Tummelplatz-Graz database will be also released for public to make further references and comparison easier.

**KEY WORDS**
Computer Vision, Semantic Segmentation, Redundancy, Multi-view segmentation

## 1. Introduction

Context-based interpretation of a scene captured in a single digital image has been addressed in several papers [2][3][6]. However, one may argue that progress is slow. This class of "hard problems" may become more tractable if one generalizes the input data to consist not of a single image, but of a stack of multiple images. We can denote this as "redundant" or "multi-view" input data. Therefore in our image databases, we usually want to employ multiple images of any given scene. Each of these images, when processed individually, will provide us with a specific interpretation. The purpose of this paper is to examine how multiple interpretations from multiple images differ and complement one another to improve the overall result. The effect of multi-view imagery on various geometric scene analyses has been established [5]. It is less well understood how the interpretation of a scene is affected by the transition from a single image to a multi-view image stack.

Our work is motivated by the need to interpret scenes as part of establishing an Internet-hosted Exabyte 3D World model [10]. The need to address the human scale of such a World model leads one to consider street side images, either via the use of an organized industrial sensor approach [4] or via crowd sourcing based on user-provided imagery [14].

To reflect both of these approaches, we have collected two initial image data bases. The first database is an Industrial System dataset. Images are taken by a calibrated multi-camera apparatus mounted on a car (see Figure 1.1).

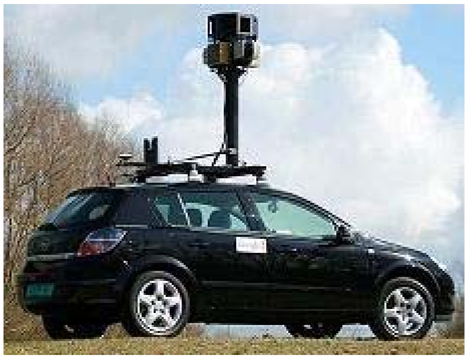

**Figure 1.1**: An example of a camera system mounted on a car. It is designed to cover wide viewing range.

This setup creates overlapping images with a rigorous and calibrated geometry from a single image-taking position, and delivering for each object point multiple images from that single sensor position. By moving the car and repeating the image collection, the level of redundancy gets further increased. Carrying along a scanning laser arrangement with the imaging sensors provides one with additional range information and means to match the images. Figure 1.2 (a) is an example of a data set that consists of 250 images from each camera on the car platform. In our work, we used the input of only two cameras – one sideways and one frontal-sideways tilted camera. The data base supports investigations into the issues of the types of redundancies, namely multiple images, all taken with parallel optical axes from different camera positions; or multiple images all taken from a single position but with different directions for the optical axes, and various hybrids between these two concepts.

The second database consists of 110 amateur photos taken with a handheld camera at the Tummelplatz in Graz (in this city's historical center). We augment the image data by manually collected ground truth: using one key frame, we manually interpret a selected collection of facades. Furthermore, the sparse 3D point cloud was created for a subset of this dataset to provide better image matching. Figure 1.2 (b) illustrates the data set. In this data set we also have sufficient images to be able to group them by similarity of their optical axes by dissimilarity due to differences in position and orientation of the optical axes, and by geometric resolution.

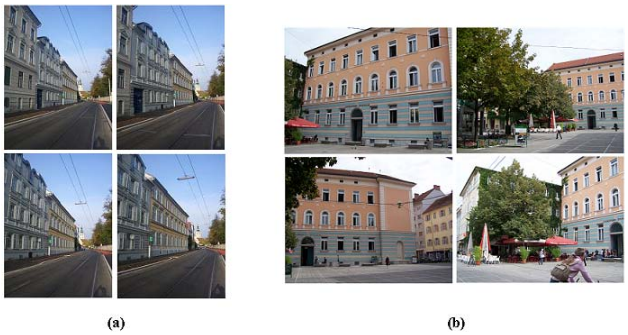

**Figure 1.2:** Example of the image type from the Industrial System database (a) with the parallel optical axes providing a high level of redundancy from sequential exposures in a moving vehicle; in this example the optical axes are pointing halfway forward. In the Tummelplatz-Graz database (b) the viewpoints and viewing directions of manually collected images can differ significantly for each object.

Being limited to just two data sets does not permit us to study the results as a function of various object types. For this to be possible, we need to increase the variation of objects and scenes being studied – to be the subject of ongoing work.

## 2. Interpretation of single image

Both datasets have been collected in the center of the city of Graz, and both contain a large assortment of objects with high interclass variety. In our semantic segmentation, we consider the following classes: sky, cloud, roof, façade, vegetation, circulation spaces, grass, shadow and unidentified. We use the workflow proposed in [13] to identify these classes in the image. Semantic segmentation applied in our approach resembles the work of Hoiem [6], but includes the idea of geometric context to achieve an improved performance.

Firstly, the image is segmented, using only visual features. We use the over-segmentation method, with small initial segments, to capture the detailed layout of the scene. Subsequently, the small segments are merged with rules to maximize the likelihood of merging only segments within the same object. The final segmentation usually provides only few large area segments in the image (two segments per building façade in average). Secondly, we proceed with the verification of classification based on geometric context. In this case we assume that some underlying spatial relations exist between the classes (see Figure 2.1 for an example). The classification hypotheses for the segments directly examine those spatial relations. For this purpose, the discriminative random fields (DRF) method [8] was implemented. DRF are a special case of Conditional Random Fields as introduced by Lafferty [9]. This modification allows us to employ discriminative classifiers instead of standard potentials. The probability distribution in this case can be expressed as:

$$P(\mathbf{x} \mid \mathbf{y}) = \frac{1}{Z}\exp\left(\sum_{i\in S} \mathrm{A}_i(x_i,\mathbf{y}) + \sum_{i\in S}\sum_{j\in N_i} \mathrm{I}_{ij}(x_i,x_j,\mathbf{y})\right) \qquad (1)$$

where $\mathrm{A}_i$ is an unary (association) potential, representing the classification of the image segment i. $\mathrm{I}_{ij}$ is a pairwise (interaction) potential, which denotes the spatial rule between the segments i and j, respectively between their assigned classes.

It is shown in [13] that this approach can improve classification mainly for those areas that have strong spatial relations, for example the pair of façade-roof. For the classification of building façades the achieved success rate was 93.7%, thus having this portion of all facade pixels actually identified as belonging to a facade.

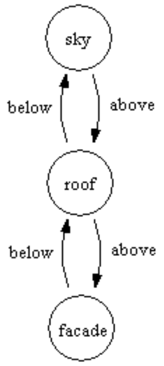

**Figure 2.1**: Example of geometric context between object classes.

In real scenes, the geometric context between segments or classes can be considered an invariant, but in a projective image of that real scene, the geometric context can be distorted. This is especially the case in composite scenes, where more than one instance of the same class is located: for example, the façade of one building can be projected above the roof of another building. Therefore, the impact of verification by context is view-dependent and can be considered as the source of discrepancies in semantic segmentations of multiple views. Our assumption is that in a redundant database, the impact of the geometric context is different for any two images of the same object that were taken from different positions, or under a different angle of view. In addition, multiple views offer the option of considering the 3rd dimension and in the process to accommodate for partial occlusions and for projection effects.

Another source of a consistent error in this type of segmentation can be observed near the border of the areas. For example, the border between the façade and the street level (ground) is often visually not well defined, as there are disturbances like shop windows, cars, pedestrians or shadows. When we consider multiple views of such an area from different angles or positions, we will get different visual information. Therefore we can assume that results of the semantic segmentation will differ for each image in the multiple-view scheme. This discrepancy can be observed mainly in the objects farther away from the camera, or in out-of-focus objects, for instance in blurred borders that provide a challenge for the segmentation. This observation allows us to assign a weight to the classification results in multiple views based on range, and range information will thus be desirable and useful.

## 3. Concept of redundancy

Content redundancy in the image database can be considered as a source on new information for the image interpretation. In this paper, we will examine several types of redundancy in regard to the position of cameras.

**(a) Multiple views from a single position with rotated optical axes**

This type of redundancy is usually present from industrial systems, where the multiple cameras are aligned in a ”star formation” (see Figure 3.1). In this case, the rotation between images is well established and calibrated; the overlap areas between images provide “redundancy” in precisely defined manner. Also, this kind of setup may occur in crowd-sourced type datasets, when a user (photographer) makes different images from one single position. The rotation parameters will not be known in this case and the redundant areas must be established through a search for correspondences in the images. There are no geometric differences for a given object, but the context may change in multi-view, as well as the visual features of the objects.

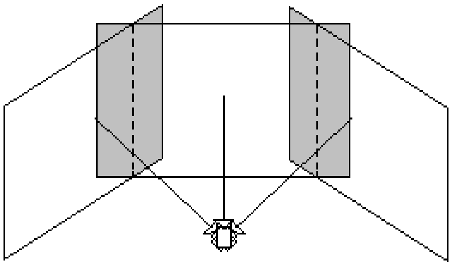

**Figure 3.1:** Camera setup in star formation. Gray areas denote redundant regions from overlaps in images taken form one single position.

**(b) Multiple views from varying positions with parallel optical axes**

This type of redundancy is generally present from systematic environment mapping (see Figure 3.2). It will result from industrial systems or from hand-held cameras if a purposeful “strip” of images is being collected, often this is the case in planning for a 3D reconstruction. The translation of the pose between the views is more or less regular, but in a natural environment, the high level of regularity is sometimes difficult to achieve.

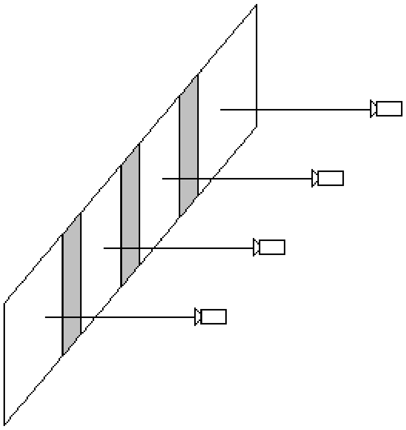

**Figure 3.2:** Camera setup with parallel axes. Gray areas denote redundant regions, typical for images acquired from a moving platform like a car.

**(c) Multiple views from varying positions and variable directions**

This type of redundancy is present in an unorganized dataset, usually obtained from hand held cameras. It can be observed in a crowd-sourced database that provides a large volume of data. The lack of organization causes the difficulty. Even the state of the art block adjustment algorithms today need dozens of views of the single object to correctly establish matches. Several approaches have been designed to create some kind of organization structure in this type of datasets. Usually, some number of correspondences has to be established first [12][1] and the camera parameters have to be determined [7]. It is therefore assumed that the overlaps and thus the “redundant information” can be obtained from this type of data.

Considerable portions of images taken in urban environment will contain temporal objects. These are located in the scene only for short periods, and are mostly people, animals, transportation vehicles, perhaps others. We do not consider them to be a specific class of objects and are therefore considering them as “undesirable” since they cause occlusion for relevant objects. In redundant databases, images of the same scene taken from different viewpoints and directions also will be taken at different times, and may support an effort to recover those occluded portions of a scene.

## 4. Image matching

In the general case of urban imaging, a block of images would be triangulated in today’s typical workflows as illustrated by Photo-tourism and Photosynth [14]. We also employed this approach and created a sparse 3D point cloud from the subset of Thummelplatz dataset. The algorithm described in [7] was used to extract this point cloud (see Figure 4.1).

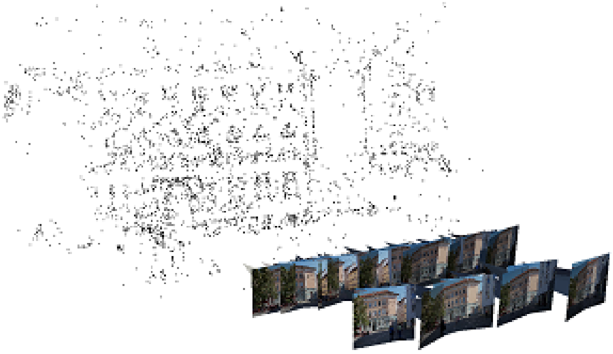

**Figure 4.1:** An example of 3D point cloud in the Tummelplatz dataset. This point cloud was created from 28 views and consists of 3498 points (thus of 125 points per image in average). Of a given façade one has 2623 points to work with.

As our goal in this paper is to match the building facades between two images pixel-by-pixel, sparse point cloud does not provide us with enough data for this. It is necessary to interpolate the positions of pixels between the points belonging to a point cloud inside the façade. We can operate with a simple assumption, that the area between two façade points is planar. In a perspective imaging, a planar object is mapped into the image plane by a projective transformation. Establishing the parameters of this transformation can provide image matching even for pixels not belonging to a point cloud. We merely need to identify at least 4 façade points in each image. We can use four non-collinear points from the point cloud, or when the point cloud is not present, we can mark these points manually and assign world coordinates for them. The perspective transformation matrix can be defined uniquely, if image and object coordinates of at least four points are measured. The relation between the point in the image plane **x** and the point in the world plane **x’** can be defined as **x’** = H**x**, where H is the projective transformation matrix. The parameters of matrix H can be computed from 4 corresponding point coordinate sets, or alternately can be derived from the certain metric properties such as length ratios and angles, as described in the work of D. Liebowitz [11].

For the purpose of testing the segmentation results, the borders and the inner area of the building façade were manually labeled. By associating with each façade in object space a unique identifier (number), we can automate the matching task for each group of images of the same façade (see Figure 4.2).

In an Industrial System database, we used the laser scanner data in similar way as a point cloud. Given that each image is geo-tagged, the position of a laser scanner point on the building façade in the world coordinate system can be projected back into each image. This will provide us with image and object coordinates of a sufficient number of object points so that the image-object relationship is fully defined

We use this simple method to relate the overlaps of the images to one-another and to then study the differences in the segmentation from image to image in the overlaps.

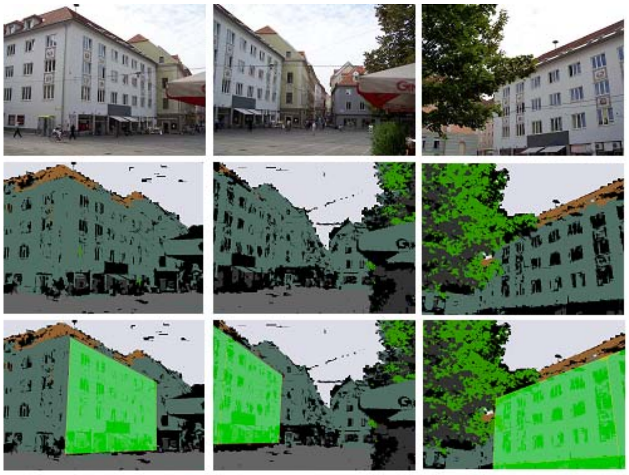

**Figure 4.2:** Image segmentation and matching. Top row – original images of the same objects from different view point. Middle row – semantic segmentation of the individual images. Dark green – façade, brown – roof, light green – vegetation, white – clouds, gray – ground. Bottom row – building façade is marked and matched.

## 5. Experiments

### 5.1 Simple application of multiple views

For the purpose of testing the multiple-view image interpretation, we have developed an annotation tool. This allows us to identify facades and the correspondences between the planar objects in the images. For each façade, the perspective distortion is computed from four non-collinear points from a point cloud (if present) or from points manually marked. The position in world coordinates for each point of the façade is computed through interpolation between three closest point cloud points. The identification number for each façade helps in automating the work with multiple images. We also identify objects that generate occlusions such as vegetation, pedestrians or cars. This type of annotation can provide us with pixel-by-pixel correspondences between the images.

For each image, a semantic segmentation is being performed as if it was all by itself, and in accordance with section 2. Our task in this experiment is to assess, how the results of the façade segmentation differ between images, and the identical object areas do get defined by means of image matching as previously discussed. The framework can be described in the following steps:

1. For each façade object, identify the group I of images, where it is located and annotated.
2. Compute the perspective transformation matrix $H_i$ for each image $i \in \mathrm{I}$
3. For each point $x_{ij} \in \mathrm{F}_{ij}$, where F is the façade in the image $i$ with the identification number $j$, transform $x_{i,j}$ into the world plane coordinates $x'_{i,j} = H_i x_{i,j}$
4. $\forall x_{ij}$ compute the new classification as $$s_{ij} = \frac{1}{Z} \sum_{i \in \mathrm{I}} w(x_{ij}).c(x_{ij})$$ where $c()$ is the classification of façade pixel $x_j$ as façade in image $I$ and $w()$ is the weight function. Z is the normalizing factor, setting $s_{ij}$ into <0,1> interval
5. Compute the new classification as a result of $s_{ij}$ for each pixel of the façade.

We designed several scenarios according to the concept of redundancy described in Section 3. Three scenarios are identified for the Tummelplatz-Graz database as follows:

a) stable position, rotated optical axes (SPRA)
b) varying position, parallel optical axes (VPPA)
c) varying position, varying axes (VPVA)

The industrial system (IS) is considered as a separate case with a varying position, parallel optical axes and a high level of redundancy.

In our first experiment, the weights $w(x_{ij})$ are set to 1 for each image. This approach was chosen to demonstrate, that even the simple summing of classification through all images can provide improved results over single image. Pixel $x_{ij}$ is then classified as a façade, if $s_{ij} > 0.5$ (see Figure 5.1).

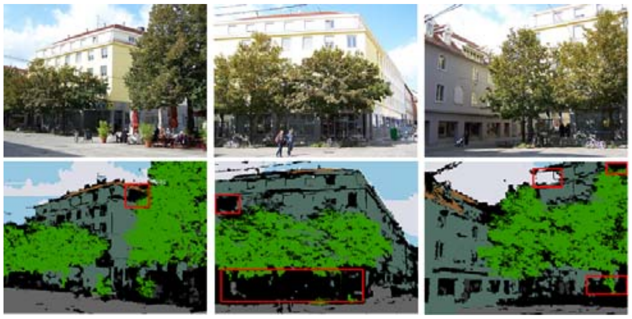

**Figure 5.1:** The segmentation of three different views of the same object. Each segmentation shows error (in a red box) different from the others.

Results from this experiment are summarized in Table 5.1. For each of the 4 overlap cases, we produce three numbers. "# img" is the number of images used in the scenario; "Single img" is an average result of classification for each image in the scenario separately. This number is expressed in a percentage of all façade pixels that were correctly classified as a façade class. The row "Multi img" is the result of multiple view approach, as described in this paper.

| Scen. | SPRA | VPPV | VPVA | IS |
|---|---|---|---|---|
| # img | 24 | 22 | 55 | 250 |
| # img/obj | 8 | 11 | 6 | 27 |
| Single img | 93.9 | 94.2 | 93.4 | 89.2 |
| Multi img | 96.2 | 96.9 | 95.7 | 93.3 |

**Table 5.1:** "# img" is the number of images; "# img/obj" is the average number of views of a given object point. "Single img" is the average value of correct classification of pixels in single image approach (in percentage); "Multi img" is the value of correct pixel classification in multiple views approach (in percentage).

From the results of this experiment we can observe, that the improvement in multiple views approach can be achieved in all examined scenarios. The single image approach has the highest error rate in the industrial system dataset. This is probably due to lower quality of the images (lower resolution and lens quality). But the improvement in multiple views approach is also higher in this scenario. It is assumed, that the high level of redundancy may be the contributing factor in this result. It is therefore assumed that this scenario can benefit the most from the redundancy in a dataset.

### 5.2 Classification consistency as a function of distance from the camera

A second experiment examines the effect of redundancy in regard to the distance of the objects from the camera. It is assumed that distant objects are more difficult to classify, as they contain larger pixels and thus less information about the object, but the relationship between the distance and the classification result is unclear. In this scenario, we use the industrial system database with laser range data to classify and match objects. We are comparing the classification of areas of building facades at different distances from the camera. The area of the façade is considered consistently classified if it is labeled as a façade, or unidentified. The results can be read from Figure 5.2. We see that at a distance of 10 meters, 95% of the façade pixels are consistently being classified as "façade" or unidentified. Going farther way to 40 m, this reduces to a level of 84%.

This result can be used in the further experiments, to derive a distance dependent weight for the classification in image overlaps or redundant databases. The classification of an object closer to the camera is at a higher confidence than that of an object that is farther away.

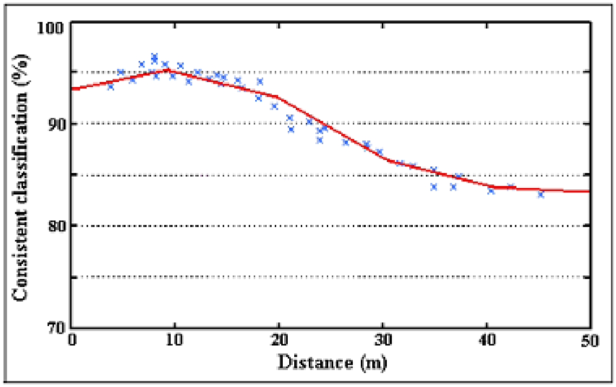

**Figure 5.2:** Relationship between the distance from the camera and the consistency of a façade classification. Values are plotted in blue for objects at various distances from the camera; the red line is an average value.

### 5.3 Multiple views classification based on distance

In this experiment, we apply our previously extracted function of distance based classification consistency to improve the algorithm described in section 5.1. We set the weight function $w(x_{ij})$ as a function of distance from a camera. This will provide the weighting for each pixel, when the distance information is available. We used the subset of Tummelplatz dataset, for which the 3D point cloud is available and the Industrial System dataset with laser range data for this experiment. The results can be observed in Table 5.2.

| Scen. | VPVA | IS |
|---|---|---|
| # img | 28 | 250 |
| Single img | 93.5 | 89.2 |
| Multi img | 96.1 | 95.7 |

**Table 5.2:** "# img" is the number of images; "Single img" is the average value of correct classification of pixels in single image approach (in percentage); "Multi img" is the value of correct pixel classification in multiple views approach using the distance as a weight function (in percentage).

In this experiment, we can conclude that selecting a more appropriate weight function $w(x_{ij})$ for the classification in multiple views scenario can add some improvement. The selection of weight function is dependent on additional data, in this case, the presence of distance information. This will require the calibration of the camera (preprocessed or automatic), or some other source of data (laser scanner, for example).

## 6. Conclusion

In this paper, we presented a study on the effect of image overlaps or redundancy on the interpretation of street side images. This work is a step in the processing of large clusters of images located in various internet databases, as well as data sets produced by methodical imaging by means of industrial systems. We demonstrate that the results of image interpretation can be improved by the application of redundancy in the various multiple views scenarios and that this improvement is available even if the approach used is rather basic.

The implication on the various crowd-sourcing datasets should be considered based on the available resources. The problem with this type of data is the non-existence of camera calibration and lack of additional localization data. However, the ongoing work on automated block adjustments algorithms (Photo-tourism, Photosynth) shows that the processing of large image databases can be performed effectively and provide sufficient information for the required image matching.

In the case of the industrial system, additional data for easy image matching are usually available. This makes the application of redundancy an attractive concept, as we can achieve improvement with little effort.

Next steps in our work are to develop a dense 3D point cloud of the objects of interest and to transform the source images into a single world coordinate system. For each XYZ-point on an object surface, we will then have a redundant pixel stack ready for submission to a multi-view classifier.

An additional result from this work is an annotated database Tummelplatz-Graz. This database contains 110 digital photographs (2576x1932 resolution), with manually labeled building façades. The system of labeling supports the study of one specific facade in several images. The database contains 16 unique building facades (including historical buildings and modern architecture), each façade shown on 3-20 images. Included in the data are the annotations of occlusion from pedestrians and vegetation. The images also encompass different lighting and weather conditions. The database is released to the public at the webpage:
http://www.icg.tugraz.at/Members/recky

**Acknowledgements:**

This work has been supported by the Austrian Science Fund (FWF) under the doctoral program Confluence of Vision and Graphics W1209

## 7. References

[1] M. Bujnak, Z. Kukelova, T. Pajdla, A general solution to the p4p problem for camera with unknown focal length. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, Anchorage, USA, June 2008; 1-8

[2] G. Csurka, F. Perronnin, A Simple High Performance Approach to Semantic Segmentation. *British Machine Vision Conference*, Leeds, UK, Sept 4, 2008; 223-229

[3] S. K. Divvala, D. Hoiem, J. H. Hays, A. A. Efros, M. Hebert, An Empirical Study of Context in Object Detection. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, June, 2009, 255-262

[4] N. Haala, M. Peter, J. Kremer, G. Hunter, Mobile LiDAR Mapping for 3D Point Cloud Collection in Urban Areas - a Performance Test. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Vol. XXXVII, ISPRS Congress 2008, Beijing, China

[5] R. Hartley, A. Zisserman, *Multiple View Geometry in Computer Vision.* 2ed, Cambridge University Press 2004 ISBN: 0521540518

[6] D. Hoeim, A. A. Efros, M. Herbert, Geometric context from a single image. *ICCV 2005. Tenth IEEE International Conference*, Vol. 1, 2005, 654–661

[7] A. Irschara, C. Zach, and H. Bischof. Towards wiki-based dense city modeling. *ICCV 2007. IEEE 11th International Conference*, 2007 1-8

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

[9] J.Lafferty, F. Pereira, Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Proc. 18th International Conf. on Machine Learning*, Morgan Kaufmann, San Francisco, CA 2001 282–289.

[10] F. Leberl, M. Gruber, 3d-Models of the Human Habitat for the Internet. *Proceedings of Visigrapp,* 2009, Lisbon, 7-15

[11] D. Liebowitz and A. Zisserman. Metric rectification for perspective images of planes. *Computer Vision and Pattern Recognition* 1998, 482-488

[12] D. Nister, An efficient solution to the five-point relative pose problem.*Pattern Analysis and Machine Intelligence*, 2004, 195-202

[13] M. Recky, F. Leberl, Semantic Segmentation of Street-Side Images. *Proceedings of the Annual OAGM Workshop.* Austrian Computer Society in OCG 2009, 271–282

[14] N. Snavely, S. M. Seitz, R. Szeliski. Photo tourism: Exploring photo collections in 3d. *ACM Transactions on Graphics (TOG)*, 2006, 835 - 846

.