Scene cut: Class-specific object detection and segmentation in 3D scenes

Jan Knopp*, Mukta Prasad†, Luc Van Gool*†
*ESAT-PSI/Visics
K.U.Leuven
Leuven, Belgium
{Jan.Knopp, Luc.VanGool}@esat.kuleuven.be

†Computer Vision Laboratory
BIWI/ETHZ
Zürich, Switzerland
mprasad@vision.ee.ethz.ch

***Abstract*—In this paper we present a method to combine the detection and segmentation of object categories from 3D scenes. In the process, we combine the top-down cues available from object detection technique of Implicit Shape Models and the bottom-up power of Markov Random Fields for the purpose of segmentation. While such approaches have been tried for the 2D image problem domain before, this is the first application of such a method in 3D. 3D scene understanding is prone to many problems different from 2D owing to problems from noise, lack of distinctive high-frequency feature information, mesh parametrization problems etc. Our method enables us to localize objects of interest for more purposeful meshing and subsequent scene understanding.**

***Keywords*-Hough Transform, Min-Cut, segmentation, detection**

I. Introduction

With 3D acquisition methods becoming widely available (*e.g*. the recent introduction of the Xbox Kinect), it is important to provide processing suites that are at par with the image processing tools available to date. A lot of effort has gone into effective 2D object class recognition, detection and segmentation despite problems like clutter, occlusion, transformations *etc*. In contrast, 3D object class detection has so far mainly focused on isolated objects. Despite artificially introduced noise and occlusion, the 3D data tends to come from one and the same object. In reality, data is often gained from an entire scene at once. The resulting point clouds or meshes have not been pre-segmented. Several 3D approaches to recognize, locate and segment specific types of objects in such data risk to fail. The goal of our work is therefore, to find and segment objects in cluttered 3D scenes.

Figure 1. **Example of the input scene.** Given a 3D scene as the ones shown here, we want to find objects of a specific class *e.g*. the bike or the woman. We learn object class models and use this to automatically detect all instances of the class in the scene.

3D reconstruction and scanning methods have matured in recent times, which enables us to process this data for further tasks. Systems like the Xbox Kinect now have the ability to retrieve 3D scans of scenes. While primarily used for the purpose of pose estimation, they also allow for more efficient methods of shape classification, recognition, retrieval and general scene understanding. Typically range-scanning or Structure-from-Motion (SfM) systems do not differentiate between the individual objects in the scene in the process of reconstruction. A raw 3D point cloud is usually the output which is converted to a mesh. The process of mesh construction is naïve and doesn't differentiate between connecting parts of the same object or splitting two different objects into separate meshes. Some methods improve on this by solving for optimal meshing that maximizes reprojection photo-consistency. Also, there is inherent noise and ambiguity in the point clouds themselves. In this context, the process of combining the process of recognition and segmentation can be useful for the purpose of scene understanding and recovery. This can help us make more complex inference for applications in *e.g*. navigation or animation. Importantly, this is useful for generic scene understanding *i.e.* the woman is sitting *on* the chair *near* the dining table with a bottle *on* it (see Fig. 1). Understanding relationships further helps in recognition, efficient shape indexing and retrieval, informed meshing *etc*. What's more, domain knowledge of the objects can be used even for complex tasks of object completion *etc*.

II. Related work

In 3D, relentless examination of feature detectors and descriptors [1], [2], [3], [4], [5] has allowed exploration of further tasks. While simple measures like curvature and

surface based similarity measures were used for object detection, invariance and occlusion can be handled far more easily with more sophisticated features. Methods such as Spin Images [2], represent features as histograms recording the spatial distribution of all points on an object *w.r.t*. a particular one. They are however sensitive to noise and have low discriminability. Adaptation of Bag-of-Features (BoF) based methods, similar to text and image retrieval have inspired shape retrieval methods such as [6]. The same BoF features are used to train classifiers for class recognition in [7].

This work draws inspiration from relatively recent research in 3D object categorization and retrieval and the more established field of corresponding problems in 2D. In 2D, a variety of works were proposed for 2D object detection ([8], [9], [10], [11]). Motivated by the problem of face detection in [8] Haar-like filters are learnt in a boosting framework to train classifiers for face detection. In order to sufficiently speed-up the process, the candidate classifiers are considered in a cascade in order of complexity. In ISM [9], a star configuration of object features with respect to the object centre is learnt during training. At test time, a forward and backward pass of voting decides the best object configuration. Pictorial structures based methods [12], [10] represent objects as a collection of parts connected by virtual springs. The probabilistic approach is employed to detect the optimal object configuration in 2D scenes. A more recent approach of Nishino and Bariya [13], treats the object recognition problem as one of finding scale-invariant correspondences between the model and the target scene. To make the problem tractable, matching is performed hierarchically to ensure that subsequent matches maintain geometric consistency with the previous ones. This enables culling the space of potential possibilities and perform efficient matching for objects. The downside is that each object must have a tree representing it, and it's not clear how well this approach would generalize for variation due to deformation. Here we are additionally interested in the ability to detect instances of an object class, in the presence of noise, clutter and deformation in a populated scene. In other words, we are interested in class-specific (than object-specific) detection, *i.e*. see Fig. 1 where the goal is to learn and find all instances of the class (person) in a given scene with noise, clutter and deformation. A focused sub-problem in this area deals with specific problems like reconstructed cities and architecture (the labelling problem of [14], [15], [16]). Information about the ground plane or several distance constraints are employed for good performance. In this work, we try to learn and employ more general shape models (ISM) without this kind of specific information.

2D image segmentation has also been a well researched field. Following the clustering based approach of normalized cuts [17], the Markov Random Field (MRF) based methods of Min-Cut [18], [19] became very popular. While the original application of Min-Cut had very simple neighborhood based consistency priors, subsequent versions applied sophisticated unaries and data-dependent pairwise terms on Conditional Random Fields (CRF) [19], [10]. The ability to easily inject higher-level object detection information in the context of class characteristics makes Min-Cut based segmentation appealing from both a theoretical and practical point of view.

There has also been increasing interest in the area of 3D segmentation. Kalogerakis *et. al*. [20] learn about segmentation of shapes and configurations and subsequently treat it as a CRF optimization problem of finding segments and their mutual configurations. Such a method can be used for prototyping, texturing *etc*. While they tackle the problem of learning seamlessly joined segments with functional roles on a single object's surface given segmented training data, we concentrate on learning to segment class-specific instances in 3D environments using the power of recognition techniques, from shape instances alone.

We combine the long-range cues of object class detection with the more local MRF based segmentation techniques for combined recognition and segmentation. The goal is to label each scene vertex by an object class (or background's) label and generate clean segmented meshes in the process. We mix object localization (using ISM [4], [9]), with Min-Cut based segmentation [18] in order to partition a mesh-based scene graph into its component objects. To the best of our knowledge, this is the first paper recognizing class-specific objects in the scene without additional constraints. We employ recognition techniques instead of hand-tuned applications (connections to background plane, user-marked segments *etc*.).

Overview: We introduce the problem of combined object detection and segmentation in § III. The ISM and Min-Cut based parts of the proposed method are described in § IV. Then, the experiments are described in § V and the results summarized in § VI.

III. Problem statement

The scene S is defined by a set of 3D vertices (acting as graph nodes) $\mathtt{V} = \left\{ \mathbf{v}_k = \begin{bmatrix} x_k \\ y_k \\ z_k \end{bmatrix} | k \in \{1 \dots N_{\mathbf{v}}\} \right\}$ connected by the set of edges $\mathtt{E} = \left\{ \mathbf{e}_l \mid l \in \{1 \dots N_{\mathbf{e}}\} \right\}$, where every edge is a pair of vertex indices $\mathbf{e}_l = \{i, j\}$. Given a 3D scene S, our goal is to locate and segment instances of an object class.

We combine the object localization technique of ISM with the segmentation technique of Graph Cuts for combined detection and segmentation. First the shape and appearance of 3D object classes is learnt in the class model $\mathcal{T}$ using ISM from training data and used to detect multiple instances of the object class in S (as shown by [9], [4], [7], [21]). Next, an energy function is specified for the purpose of segmentation in a conditional random field (CRF) frame-

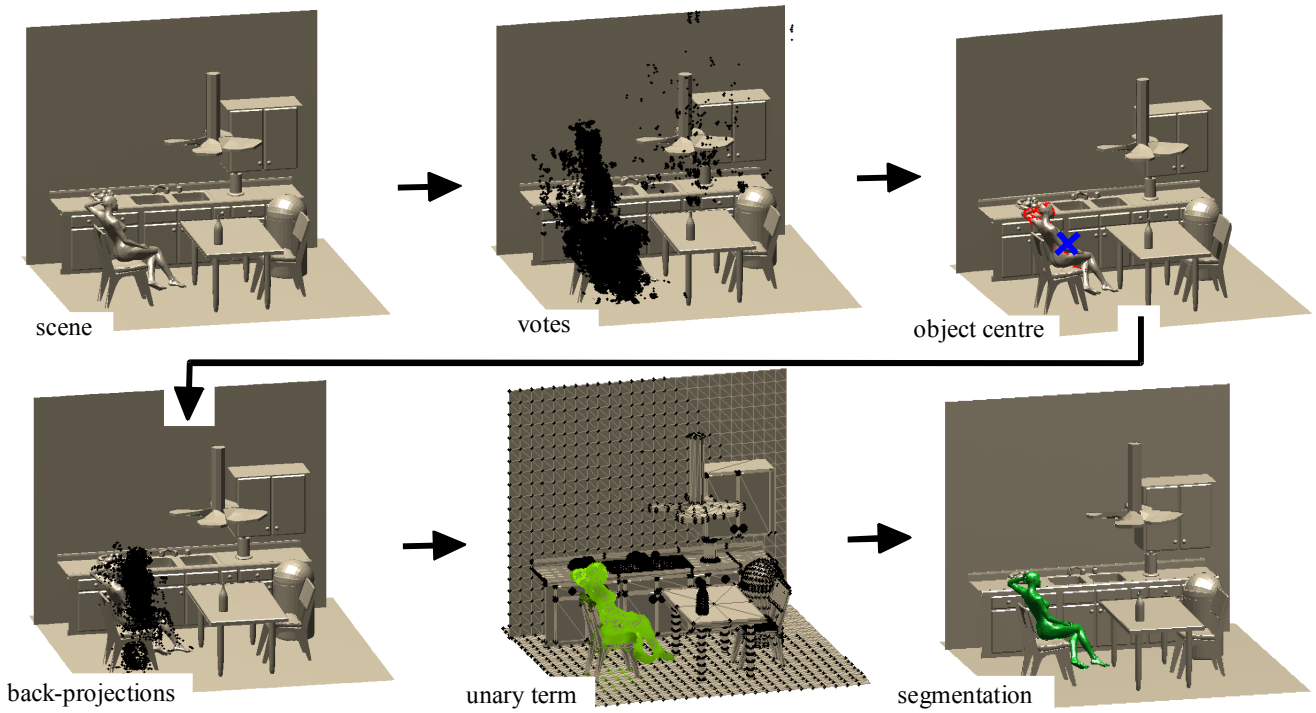


Figure 2. **Using ISM to locate and segment the *woman*:** (a) A complex scene containing the class of interest:*woman*, is given. (b) Votes for the object centre, are cast (in black) from visual word occurrences corresponding to the closest visual word for each vertex feature. (c) The location of maximum vote density is identified as object centre (blue). We also plot the correctly contributing vertices (in red). (d) The back-projection of visual word occurrences from the estimated object centre are shown. The unary term is derived from a Chamfer-like distance to these, defining the unary term (plotted in green). (e) The *woman* is segmented (in green) by performing Min-Cut with this ISM unary and Ising prior.

work and efficiently optimizing using Min-Cut, for effective segmentation in the presence of clutter, occlusion and noise.

IV. The Method: ISM and Min-Cut for Object Detection

In this section, the proposed approach is discussed. We first give a short overview of ISM [9], [4], [21] for object detection in § IV-A. Then, in § IV-B we formulate the segmentation task as a graph optimization problem. The individual terms constituting the objective are defined in § IV-C, § IV-D and § IV-E.

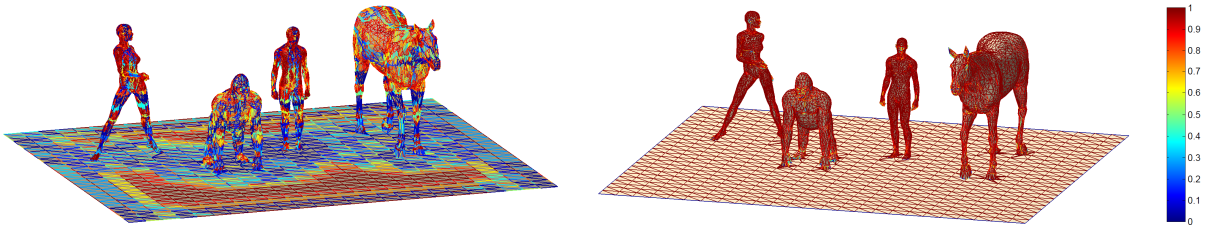
A. ISM for object localization

Hough Transform based methods, such as ISM, enable detection of an object class instance in a two step forward-backward voting scheme. Given a set of training shapes $\{S\}_{1...\mathtt{T}}$ of a class, denoted as in § III, we aim to learn a class model $\mathcal{T}$. In addition to vertex position $\mathbf{v}_k$, each vertex node of a shape S is also associated with a description of its surface properties in the form of scale σ_k, a shape descriptor $\mathbf{d}_k$ (*e.g.* HKS [5], SI [2], 3DSURF [4] *etc.*), relative position of object centre $\boldsymbol{\lambda}_k$ *etc*. K-means clustering is performed on the descriptors of all vertices, shapes and classes together. K cluster centres ($= 10\%$ of the total number of descriptors) are obtained, which form the visual words of the vocabulary $\mathtt{W} = \{\mathbf{w}_i \mid i = 1...K\}$, forming the basis for the traditional bag-of-words approach [22], [6]. Each training feature is assigned to the closest visual word if within threshold distance $\Delta(= 0.08)$ in the descriptor space as in [9], [4]. For each visual word $\mathbf{w}_k$, the set of M_k visual word occurrences (vertices assigned) and all corresponding information: $\sigma, \lambda, \mathtt{D}$ *etc*., is stored. The ISM class model is therefore: $\mathcal{T} = \{\{S\}_{1...T}, \sigma, \lambda, \mathtt{D}, \mathtt{W}, \Delta, K, M\}$. At test time, a test scene vertex $\mathbf{v}_i$ is assigned to the closest visual word $\mathbf{w}_k$. The cluster's training visual word occurrences generate a set of votes $(\boldsymbol{\lambda}_i | i \in \{1 \dots M_k\})$ for their estimate of the object centre $\mathbf{O}$. The location of maximum density of such votes defines the estimated centre $\mathbf{O}$ of the object. The features that contribute to the final estimated $\mathbf{O}$ may be re-traced to estimate the object extent.

B. Min-Cut for efficient object segmentation

For a scene graph $S = (\mathtt{V}, \mathtt{E})$[1] as defined in § III, the scene segmentation problem can be posed as a labelling task

[1]The graph $S = \{\mathtt{V}, \mathtt{E}\}$ is easily obtained from a triangulated mesh and helps express S in a graph framework. Meshing in point clouds usually employs spatial proximity to draw edges between nodes in a graph [16]. Thus, the process is intertwined with subsequent graph processing. In this work, we assume graph based representation does not result in loss of significant generality or information.



(a) Vertex similarity (b) Plane-like surface

Figure 3. **Data-dependent pairwise terms:** (a) The vertex similarity measure § IV-E1 and (b) the plane-like measure § IV-E2, are visualized by the texture map of the above scene. **Red** color represents stronger edges with high penalty of breakage (blue represents weak edges prone to being cut). In (b) the blue edges are mostly at the point of contact of these models with the ground.

of whether a scene vertex (graph node) $\mathbf{v}_k$ belongs to an object ($f_k = 1$) or background ($f_k = 0$). The goal is to find the binary valued label vector $\mathbf{f} \in \{0,1\}^{N_\mathbf{v}}$. An ideal segmentation would ensure that: (i) the detected foreground segment resembles obeys the object class characteristics, and (ii) the foreground and background segments are smooth while accounting for the observed data. *E.g.* smooth areas of the mesh should be more likely to belong to the same segment, areas on either side of a crease or dissimilar shape regions should be penalized from belonging to the same segment. The desiderata, captured in terms of the unaries and pairwise terms of a conditional distribution, allows the problem to be formulated as: find the optimal $\mathbf{f}$ that maximizes the distribution $\text{argmax}_f\, p(\mathbf{f}|S,\theta)$, for S denoting the current test scene and θ denoting a set of parameters for the graph (including the learned ISM class model $\mathcal{T}$). This can be expanded as:

$$p(\mathbf{f}|S,\theta) = \frac{1}{Z(S,\theta)} \exp\Big(-Q(\mathbf{f};S,\theta)\Big), \tag{1}$$

$$Q(\mathbf{f};S,\theta) = \sum_{i=1}^{N_v} \Bigg(\underbrace{\phi(\mathbf{v}_k;f_k,\theta)}_{\text{unary}} + \sum_{j|(i,j)\in\mathtt{E}} \Big(\underbrace{\Lambda_\psi \cdot \psi(f_i,f_j;\theta)}_{\text{prior}} + \underbrace{\Lambda_\chi \cdot \chi(f_i,f_j,d_i,d_j;\theta)}_{\text{data-dep}} \Big) \Bigg) \tag{2}$$

where $Z(S,\theta)$ is the partition function. The **unary** potential ϕ represents the data likelihood. The second, **pairwise prior term** ψ (weighted by Λ_ψ) encourages neighbouring vertices to have similar labels. This imposes a generic smoothness on the segmentation. The third, **data-dependent pairwise term** χ (weighted by Λ_χ) encourages nodes with similar characteristics, such as shape (encapsulated in the shape descriptors $\mathtt{D}$), to have the same label. χ differs from ψ, as it depends on the observed data. The set of parameters for segmentation (including ISM parameters $\mathcal{T}$) are included in the set: $\theta = \{\Lambda_\psi, \Lambda_\chi, \alpha, \mu, \beta, \kappa, \gamma, \rho, \mathcal{T}\}$ and are defined in greater detail. The energy function Q can be solved efficiently using Min-Cut [18], [23], [10]. The energy terms are described in greater detail now. focal independent variable

C. The detection-based unary term

The areas more likely to be a part of the object in the ISM detection step § IV-A should be more likely to belong to the foreground segment. In the forward voting phase of ISM, the centre of a detected object(s) $\mathbf{O}$ is estimated (see Fig. 2) as the location of maximum density in the voting-space. The hypothetical visual word occurrence locations can be estimated by back-projecting from $\mathbf{O}$: $\mathcal{B} = \{\mathbf{b} = -\boldsymbol{\lambda} + \mathbf{O}\}$ (see Fig. 2), where λ was a visual word occurrence based vote for $\mathbf{O}$. When the vote is correctly cast, the estimated visual word occurrence location agrees with the scene vertex casting the vote. The unary data-likelihood for every scene vertex $\mathbf{v}_k$ is therefore expressed as the weighted mean distance between the vertex and the ρ closest back-projections:

$$\phi\left(\mathbf{v}_k\ ; f_k, \theta\right) = \delta_{f_k,1}\varphi(v_k; f_k, \theta) + \delta_{f_k,0}\left(\kappa - \varphi(v_k; f_k, \theta)\right), \tag{3}$$

$$\varphi\big(v_k; f_k, \theta\big) = \frac{1}{\rho}\sum_{i=1}^{\rho} \exp\left(\frac{-||\mathbf{v}_k - \mathbf{b}_k^i||^2}{\alpha^2}\right) + f_c\big(v_k\big), \tag{4}$$

where $\mathbf{b}_k^i$ is the i^{th} closest backprojection for the v_k vertex. Empirically, $\rho = 10$ is found to be effective and $f_c\big(v_k\big) = 1$, if v_k correctly voted for the estimated $\mathbf{O}$, otherwise 0, encouraging correctly voting vertices to form foreground. Note: Visual word occurrences in wrong configuration are penalized in this unary term. The parameter α was found to be optimal when set to 5.7% of the model scale.

D. Pairwise smoothness prior

The Ising prior ψ: Neighbouring vertices i, j are encouraged to have the same label with the following prior ($\gamma = 1$ here):

$$\psi(i,j) = (1 - \delta_{i,j}) \cdot \gamma. \tag{5}$$

E. Data-dependent pairwise terms

Data-dependent pairwise terms can be used to further enhance the segmentation, two ways of which are now described.

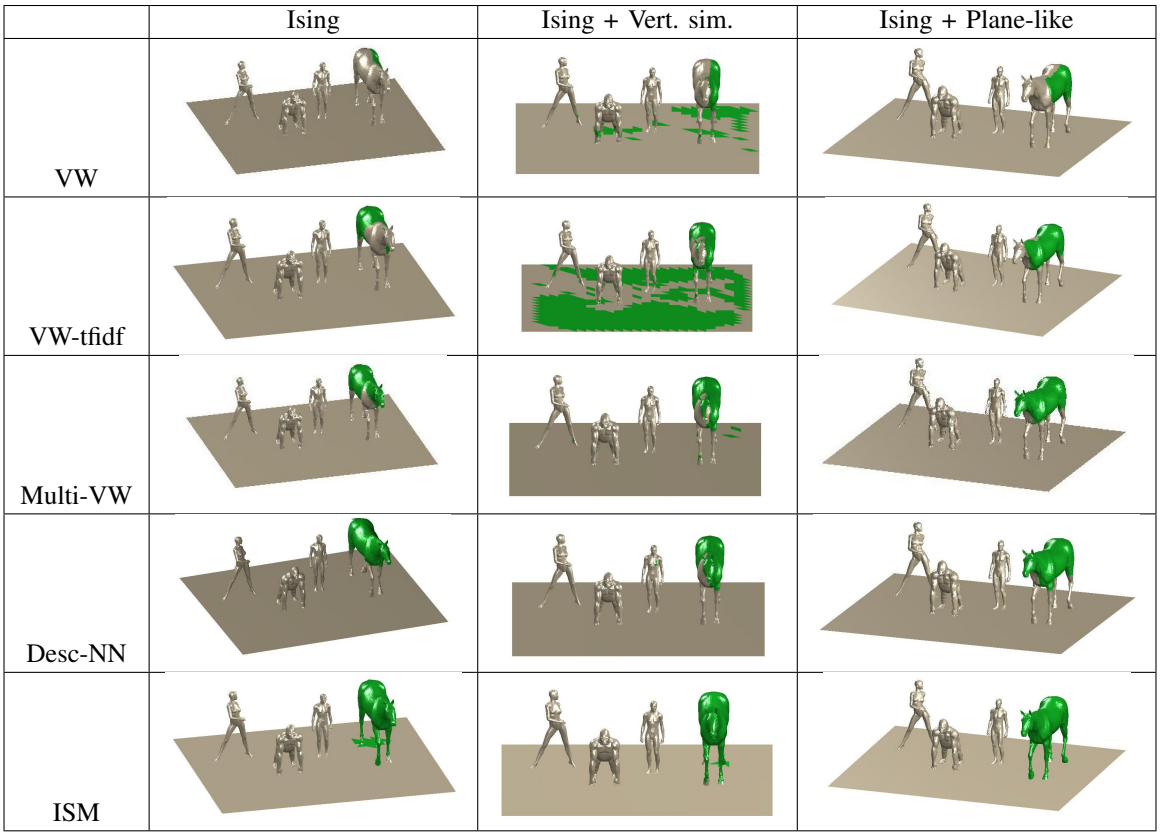


Figure 4. **Segmentation with different unary and pairwise terms:** Detection and segmentation is run for the horse class with weights $\Lambda_\psi = \Lambda_\chi = 0.77$ (empirically set), for all methods for this scene. Note that ISM (best performance with the plane-like pairwise term) performs optimal segmentation here.

1) Vertex similarity: Neighbouring vertices i, j with similar shape characteristics (captured by descriptors $\mathbf{d}_i, \mathbf{d}_j$), are more likely to have similar labels than those that are different. This pairwise energy term is defined as:

$$\chi_v(i,j) = \exp\left(\frac{-\|\mathbf{d}_i - \mathbf{d}_j\|^2}{\beta^2}\right). \quad (6)$$

2) Plane-like surface: If an edge lies on a crease on the surface, it should be more likely to broken during segmentation. The smoothness of an edge can be measured by the change in the surface normal across the edge and this pairwise cost is summarized as the normalized dot product of the face normals:

$$\chi_n(i,j) = \frac{\mathbf{n}_{l1} \cdot \mathbf{n}_{l2}}{\|\mathbf{n}_{l1}\| \, \|\mathbf{n}_{l2}\|}, \quad (7)$$

where $\mathbf{n}_{l1}$ and $\mathbf{n}_{l2}$ are normals of faces on either side of the edge $\mathbf{e}_l = \{i, j\}$.

V. Experiments

We compare the different energy terms for the purpose of segmentation. The competing data-dependent pairwise terms have been described in § IV-E. We now define the competing unary terms in § V-A. The experimental setup (parameters, datasets, descriptors) are discussed in § V-B. Experiments on synthetic (§ V-C) and real world scenes (§ V-D) are then detailed.

A. Unary term competitors

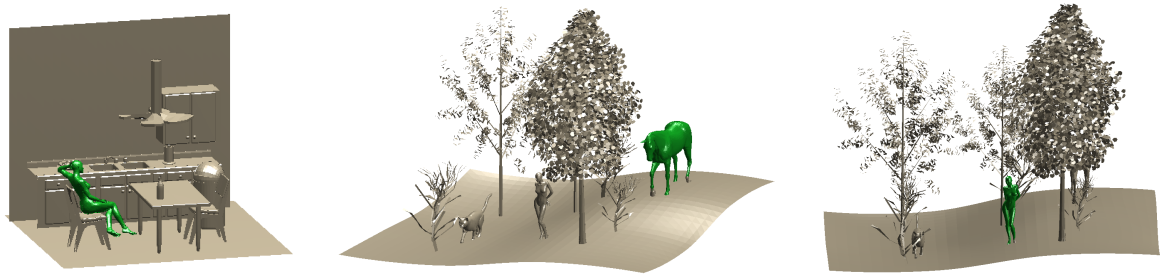
In addition to (3), we propose and evaluate other unaries for 3D detection/segmentation.

1) VW: A bag-of-feature (BoF) histogram is computed by quantizing training features to the bins corresponding to their closest visual words. The normalized histogram $\underset{K\times 1}{\mathbf{h}}$ defines occurrence frequencies of each visual word $\mathbf{w}_k$. For a test scene S, φ for a simple look-up based unary term (3) is defined as:

$$\varphi(\mathbf{v}_i; f_i, \theta) = h_{k^*} | k^* = \underset{k}{\operatorname{argmin}} \|\mathbf{d}_i - \mathbf{w}_k\|^2 \quad (8)$$

2) VW-tfidf: When learning across many classes (even though detection and segmentation may be performed for one class at one time), we want to give preference to visual words unique to the current class. Therefore the BoF histogram of § V-A1 are normalized across classes in the style of Tf-Idf [22] (see paper for more detail).

3) Multi-VW: This is similar to V-A1. The difference is in the way the BoF histograms $\mathbf{h}_k$are constructed. Instead of assigning a feature to the closest visual word during training, in this case it is assigned to all visual words closer than the



(a) (b) (c)

Figure 5. **Scene segmentation results.** Several scenes merged with objects from the TOSCA dataset which were not used in training stage. (a) detection of the person class; (b) horse class; (c) and again person class.

Figure 6. **Examples of training shapes.** 7 out of 10 shapes used to train the class of bikes.

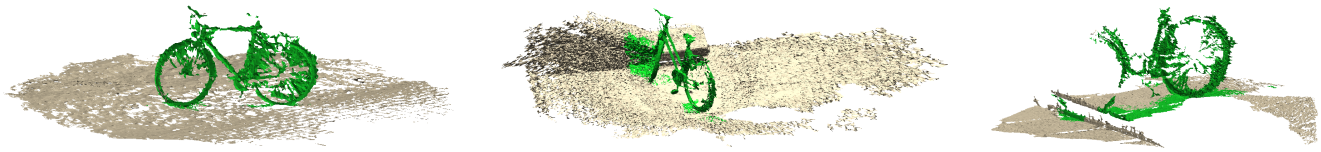


Figure 7. **Segmentation results.** Scenes were obtained using SfM algorithm from the set of camera-shots around the object of interest.

threshold Δ. φ and the unary ϕ continues to be defined as presented in (8).

4) Desc-NN: The method is motivated by Boiman *et. al.* [24], where the unary does not rely on the quantization by visual words any more and does not include supervised learning of classes. Instead a nearest-neighbour approach is adopted. Given the scene S, the unary term for the scene's vertex v_i is defined as the weighted distance of $\mathbf{d}_i$ to the closest descriptor $\{\mathbf{d}\}^{\min}$ from the entire training set of features for the class,

$$\varphi(v_i; f_i, \theta) = exp\left(\frac{-||\mathbf{d}_i - \mathbf{d}^{\min}||}{\mu^2}\right), \qquad (9)$$

Here $\mu = 0.17$ was empirically found to be optimal.

B. Setup

1) Parameters: Most parameters within $\theta, \mathcal{T}$ were found by cross-validation or empirical estimation and have been specified. The relative weights of the unary, pairwise prior and data-dependent pairwise terms were set by cross-validation.

2) Shape Descriptor: In our experiments we used modified 3D SURF [3], [4] computed at each vertex with the constant scale. This was found to perform better than the HKS [5] (experiments are not shown here)

3) Datasets:

(a) Half of the TOSCA dataset [25] is used training. For testing, 4 query scenes (see Figs. 9,8 and 5(b,c)) were manually constructed using several different (test set) objects. All test-set object meshes were fused together to create the TOSCA test scenes.

(b) Additionally, we consider a kitchen scene from the 3D lighting contest [26]. A test shape from the female class of TOSCA, was added to the scene[2] (see Fig. 5(a)).

(c) To investigate the performance on real-life scenes, 3D scene meshes reconstructed using SfM (see Arc3D [27]), are employed for the class: bicycle. Some automatic mesh simplification and downscaling is performed to obtain usable meshes. For the particular case of SfM data, 16 scenes were used, of which 10 manually segmented scenes were employed for training. This is the only input given by the user. The training data has $\sim$15K vertices/model in average, see in Fig. 6. The remaining scene reconstructions are used for testing.

C. Evaluation on synthetic data

In this section, we apply the presented methods to segment class specific objects in scenes for competing unaries (different detection strategies) and pairwise (differing segmentation regularization) terms. Note, the simple pairwise Ising prior is applied for all cases in addition to the data-dependent terms. In the training phase, the TOSCA class models—cat, man (david,michael), woman (victoria), dog, gorilla, horse, lioness and seahorse—are learnt from the TOSCA training set (see § V-B3(a). During the detection/segmentation phase, we assume the class of interest is already known. This can also be automatically performed as shown in [4].

[2]All data used in this paper are available at https://securewww.esat.kuleuven.be/psi/visics/research/test-data/.

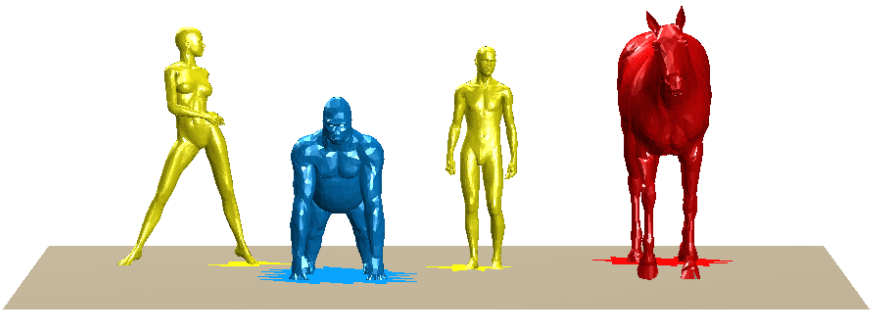


Figure 8. **Scene segmentation for multiple classes:** The ISM based unary term is used solely with the Ising prior (no data-dependent pairwise term) for segmentation with man, woman, gorilla and horse, one at a time, with weights $\Lambda_\psi = \Lambda_\chi = 0.77$. The results of each are plotted together; **Yellow** for person (man and woman), **blue** for gorilla, and **red** for the horse.

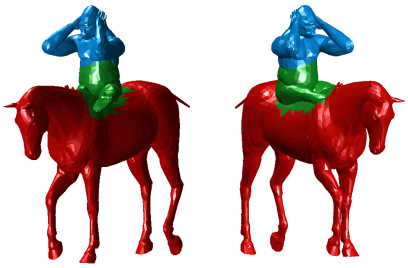


Figure 9. **Multi-class segmentation of complex scenes:** The example shows two different classes: horse and gorilla. Performing detection and segmentation of each individually, results in the green region getting assigned to both classes.

The various combinations of unaries from § V-A and data-dependent pairwise terms § IV-E are applied to the detection and segmentation pipeline. The results on the test TOSCA scenes of § V-B3(a) are shown in Fig. 4. Note that the full-fledged ISM has significant advantage owing to superior object class modelling. Competing methods of unaries suffer because they are based on simple statistical measures without accounting for the full object class structure. Among the data-dependent pairwise terms, the plane-like prior of § IV-E2 performs best.

Fig. 5 shows more results on scenes from § V-B3(a,b). Our experiments qualitatively demonstrate the success of the proposed method for combined class-specific detection and segmentation in cluttered scenes.

D. Application on real-life SfM data

Reconstruction of real-life scenes (PhotoTourism [28], Arc3D [27]) and object localization/segmentation in them [16] is an important research task that is increasingly gaining practical relevance. It has potential for great impact on applications associated with driving assistance and robot navigation systems. In nascent experiments, we attempt object class detection and localization on cluttered and noisy real-life scenes.

Qualitative results are shown in Fig. 7. While shapes are extremely noisy, full of holes and incomplete, the combination of ISM (bike class model) and Min-Cut (segmentation algorithm) correctly locate and segment the bike from background. Our results show promise for further research along these lines.

E. Future Work: multiple class localization and segmentation

We have seen how to perform detection and segmentation for a scene, when the class of interest is known and learnt. What if we want to do this for more classes simultaneously? For simpler cases such as the TOSCA test scene of Fig. 8, where the object class instances man, woman, horse and gorilla are separated, each class instance only needs differentiating and segmenting from the background. In this case, our approach is performed repeatedly for each class to give the successful result of Fig. 8. However, when object classes interact more closely, this proves to be difficult. The result of detecting gorilla (blue) and horse (red) in Fig. 9 results in overlap of regions claimed by the two classes (in green). Detection and segmentation of multiple classes within a scene therefore, is more challenging.

We present preliminary results here but the principled tackling of the multi-class problem is important from both the research and practical perspectives. In the detection phase, the classification would need 'one *vs.* all' or 'all *vs.* all' (directed acyclic graph) frameworks, for discrimination between different possible classes (including background). In the segmentation phase, this would need to be appropriately injected into the energy terms. In particular, the pairwise terms must be defined for every possible pair of possible label transitions while ensuring sub-modularity. Ensuring this can be efficiently and accurately optimized, would make this deploy-able for large scale retrieval and scene understanding problems requiring automation.

VI. Conclusion

In this paper, we have shown an approach for automatic detection and segmentation of class-specific objects in 3D scenes. Our contribution is combining the two tasks by fusing the high-level information provided by an object

detection algorithm (ISM) with the information in low-level scene priors (and data-dependent priors). As opposed to traditional voting-propagation based segmentation in the original ISM work [9], this allows for the more efficient Min-Cut to be used for segmentation. A vast amount of research in the 2D image community has employed joint recognition and segmentation of objects. This naturally inspires the use of similar techniques in 3D, but extending to 3D is not trivial. Problem-specific issues of clutter, occlusion, noise, lack of distinctive features must be handled and domain-specific priors must be exploited. Here we show how ISM based detection and Min-Cut based segmentation are combined for object class localization and segmentation in challenging synthetic and real-world SfM scenes. There are many areas ripe for the pursuit of further research, object parametrization and mesh resolution are two such areas. While we assume the mesh is sensible enough to determine a sensible solution, this issue is worth more exploration.

Acknowledgment. We gratefully acknowledge the support of EC Integrated Project 3D-Coform.

REFERENCES

[1] Saupe, D., Vranic, D.V.: 3d model retrieval with spherical harmonics and moments. In: Proceedings of the 23rd DAGM-Symposium on Pattern Recognition, London, UK, Springer-Verlag (2001) 392–397

[2] Johnson, A.E., Hebert, M.: Using spin images for efficient object recognition in cluttered 3d scenes. IEEE PAMI **21** (1999) 433–449

[3] Willems, G., Tuytelaars, T., Van Gool, L.: An efficient dense and scale-invariant spatio-temporal interest point detector. In: ECCV. (2008) 650–663

[4] Knopp, J., Prasad, M., Willems, G., Timofte, R., Van Gool, L.: Hough transform and 3d surf for robust three dimensional classification. In: ECCV. (2010)

[5] Sun, J., Ovsjanikov, M., Guibas, L.: A concise and provably informative multi-scale signature based on heat diffusion. In: SGP. (2009) 1383–1392

[6] Ovsjanikov, M., Bronstein, A., Bronstein, M., Guibas, L.: Shape google: a computer vision approach to isometry invariant shape retrieval. In: NORDIA. (2009)

[7] Toldo, R., Castellani, U., Fusiello, A.: A bag of words approach for 3d object categorization. In: MIRAGE. (2009)

[8] Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: CVPR. (2001)

[9] Leibe, B., Leonardis, A., Schiele, B.: Robust object detection with interleaved categorization and segmentation. IJCV **77** (2008) 259–289

[10] Kumar, M.P., Torr, P.H.S., Zisserman, A.: OBJ CUT. In: CVPR. (2005)

[11] Shotton, J., Blake, A., Cipolla, R.: Contour-based learning for object detection. In: ICCV. (2005)

[12] Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient matching of pictorial structures. In: CVPR. (2000)

[13] Bariya, P., Nishino, K.: Scale-hierarchical 3d object recognition in cluttered scenes. In: CVPR. (2010)

[14] Munoz, D., Vandapel, N., Hebert, M.: Directional associative markov network for 3-d point cloud classification. In: International Symposium on 3-D Data Processing, Visualization, and Transmission (3DPVT). (2008)

[15] Brostow, G.J., Shotton, J., Fauqueur, J., Cipolla, R.: Segmentation and recognition using structure from motion point clouds. In: ECCV (1). (2008)

[16] Golovinskiy, A., Kim, V.G., Funkhouser, T.: Shape-based recognition of 3d point clouds in urban environments. In: ICCV. (2009)

[17] Shi, J., Malik, J.: Normalized cuts and image segmentation. IEEE PAMI (1997)

[18] Boykov, Y., Kolmogorov, V.: An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. In: IEEE PAMI. (2004)

[19] Rother, C., Kolmogorov, V., Blake, A.: "grabcut": interactive foreground extraction using iterated graph cuts. ACM Trans. Graph. **23** (2004)

[20] Kalogerakis, E., Hertzmann, A., Singh, K.: Learning 3D Mesh Segmentation and Labeling. ACM Transactions on Graphics **29** (2010)

[21] Salti, S., Tombari, F., Stefano, L.D.: On the use of implicit shape models for recognition of object categories in 3d data. In: ACCV. (2010)

[22] Sivic, J., Zisserman, A.: Video Google: A text retrieval approach to object matching in videos. In: ICCV. (2003)

[23] Kolmogorov, V., Zabih, R.: What energy functions can be minimized via graph cuts? IEEE PAMI (2004)

[24] Boiman, O., Shechtman, E., Irani, M.: In defense of nearest-neighbor based image classification. In: CVPR. (2008)

[25] Bronstein, A., Bronstein, M., Kimmel, R.: Numerical Geometry of Non-Rigid Shapes. Springer Publishing Company, Incorporated (2008)

[26] 3dRender.com: (Lighting challenges) http://www.3drender.com/challenges/.

[27] Vergauwen, M., Gool, L.V.: Web-based 3d reconstruction service. Mach. Vision Appl. **17** (2006) 411–426

[28] Snavely, N., Seitz, S.M., Szeliski, R.: Photo tourism: Exploring photo collections in 3d. In: SIGGRAPH Conference Proceedings. (2006) 835–846