Multi-Resolution Binary Shape Tree for Efficient 2D Clustering

Csaba Beleznai, Andreas Zweng
AIT Austrian Inst. of Technology
Vienna, Austria
csaba.beleznai@ait.ac.at

Thomas Netousek
eMedia Monitor GmbH
Vienna, Austria

Josef Alois Birchbauer
Video Analytics
Corporate Technology
Siemens AG Österreich, Austria

Abstract

The analysis of discrete two-dimensional distributions is a relevant task in computer vision, since many intermediate representations are generated in form of a two-dimensional map. Probabilistic inference or the response of discriminative classification often yield multi-modal distributions in form of 2D digital images, where the accurate and computationally efficient delineation of structures with varying attributes such as scale, orientation and shape represents a challenge. The simplest example is non-maximum suppression, where typically the response of a center-surround structural element applied as a filter is used to suppress spurious detection responses. In this paper we propose a simple scheme which is capable to delineate the shape of arbitrary distributions around a local density maximum driven by a local binary shape model, resulting in consistent object hypotheses. We employ a coarse-to-fine analysis scheme where learned binary shapes of increasing resolution guide a shape matching process. We demonstrate applicability for delineating compact clusters in a noisy probabilistic occupancy map, and the capability for detecting structurally consistent line structures in a text detector response map. Results are compared to other spatial grouping schemes and obtained results demonstrate a fast and accurate delineation performance.

1. Introduction

Intermediate probabilistic representations are often two-dimensional distributions in computer vision, calling for the key task of local grouping in order to generate consistent object window hypotheses. Such distributions typically exhibit marked spatial correlation for nearby pixels implying that a local structure is present. In case of weak detection responses, background clutter and noise, knowledge associated with prior structural information - such as the expected shape of the local distribution - can help to recover weak evidence. The challenge arises from the fact that the distributions to be analyzed are multi-structured: multiple spatially extended patterns might exist at different locations, orientations and scales, thus rendering a robust and computationally efficient clustering and spatial delineation a challenge. Simple approaches refuge to binarization, where the generated segmentation result greatly depends on sensitive parameters and the presence of other nearby structures and noise. Mode seeking techniques such as the Mean Shift [4] and the scale-adaptive CamShift [2] algorithms locate the center and boundaries of a local distribution non-parametrically, but they are sensitive to noise-contaminated data and multiple nearby modes.

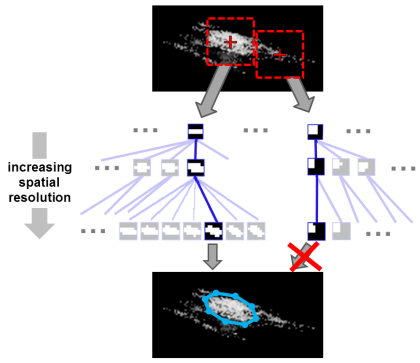

Figure 1. Our proposed clustering method analyzes a distribution (top) at mode candidate locations by retrieving the best-matching binary shape from a learned coarse-to-fine structured codebook of local structure entries. Local noise will likely match a structure pointing to a shape not centered on a sought distribution. Such candidates are discarded, while structurally validated modes are outlined (bottom) using the corresponding shape representation stored in the codebook.

In this paper we propose a novel cluster delineation scheme which employs prior, structure-specific knowledge in form of a coarse-to-fine organized binary shape codebook to evaluate the local distribution around a density maximum in terms of a best matching shape (see Figure 1). The technique is applicable for compact local distributions as well as for distributions exhibiting multiple line structures embedded into noisy data. Learning and evaluating shapes are based on a simple subdivision and quantization scheme, where the use of a hierarchical representation and integral images for mode seeking and shape evaluation result in an fast run-time performance.

The paper is structured as follows: First, Section 2 gives a concise overview on related work. Section 3 presents the proposed methodology in detail. Experimental validation and discussion are presented in Section 4. Finally, Section 5 concludes the paper.

2. Related work

Generating object hypotheses by delineating consistent maxima in a probabilistic response map is a relevant task in computer vision. Often, a **weakly constrained structural prior** in form of a local maximum search is employed. Non-maximum suppression is a common way to address this task, whether using a greedy maximum search scheme [11] or posing the search as a clustering problem [13]. Mean Shift [4] and its scale adaptive variants [2], [1] represent a similar strategy as they are simple non-parametric ways to locate density maxima and to characterize their spatial extent by determining the so-called basin of attraction [4] or estimating the local covariance. Although they are applicable to arbitrary shaped distributions and fast to compute [7], noise and multiple nearby modes tend to hinder convergence to certain mode locations.

Structure information on the expected shape of a local distribution typically brings added robustness for noise contaminated data. Linear structure detection is proposed in [14] using oriented line segments. In this context, also the use of structure-encoding templates relates to our work, such as the hierarchical contour templates proposed by [6]. Local structural elements such as *bricks* [8] and *shapelets* [3] can be learned in order to represent and analyze the local intensity and color distribution in images, and used as local descriptors for the visual object recognition task. The *Implicit Shape Model* [10] is a popular computational scheme to use local patch-based appearance [10] or binary structure [12] along with a probabilistic voting step to determine the center and the boundary of object hypotheses. Recently, structured random forests are used to estimate the edge structure [5] or the spatial distribution of semantic labels [9] present within local image patches. In these cases a learned set of binary split functions encodes structural information and yields a structured output for a test image patch.

Our approach targets the delineation task of noisy structures within two-dimensional distributions, while exploiting the idea of patch-based structure representation. Accordingly, our technique could be denoted as a structure-aware non-maximum suppression tightly coupled with mode seeking, while the above techniques focus more on image segmentation and recognition aspects using intensity information.

3. Methodology

In this section we describe the annotation, learning (Figure 2) and delineation steps of our approach in detail. We provide details for compact (denoted as *cluster*) and linear (denoted as *line*) shapes, as the delineation steps slightly differ for these two shape types.

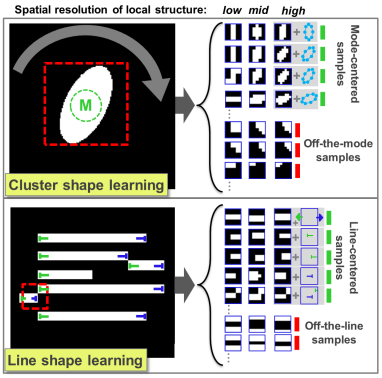

Figure 2. Illustration explaining the learning step of local structures for elliptic clusters (top) and multi-line structures (bottom). Both input images on the left are sampled (red dashed rectangle) at multiple locations and orientations (top), giving rise to a large number of local binary structures. Structures at the highest spatial resolution are also annotated with the generating elliptic contour (for clustering) or line ending locations (for line structures). Structures centered on the modes or line structures are flagged as valid (green bars).

Annotated data: For both shape types a binary mask derived from manual annotation or from synthetic data is needed. For compact *clusters* we generate an elliptical shape image - since it best represents our use case data in the experimental section - with a varying orientation (Figure 2 top). Nevertheless, a more rich shape variation could be also modeled by varying ellipse parameters (size, aspect ratio) or other compact shape types such as filled polygons could be also easily incorporated. The central region or mode of the ellipse (denoted by M in Figure 2) within a tolerance radius r_m (allowing for some spatial jitter around a mode) is set parametrically. The definition of this region will become relevant during the spatial sampling step, where structures originating from around the mode are labeled differently than shape samples from the peripheral regions. For each ellipse pose we also generate a coarse polygonal contour representation $\mathbf{v} = [\mathbf{v_1} \dots \mathbf{v_n}]$ where $\mathbf{v_i} \in \mathbb{R}^2$ denotes the position of the vertices. For *line* structures manually annotated horizontal lines serve as a basis for generating a binary image, where the left and right end points $\mathbf{t} = [\mathbf{lt}, \mathbf{rt}]$ and the line height h define a line segment.

Shape learning: In the learning step, first we derive a pool of local binary shapes at multiple resolutions from the annotated data. In order to perform a local analysis, we define an analysis window W of a size H which is used to sample the annotated binary shapes at multiple locations. The window size H is set such that the entire ellipse is captured, while for *lines* H is set to be slightly larger than the line height (see Figure 2). During sampling certain windows are centered on the binary structure (within the region M), which

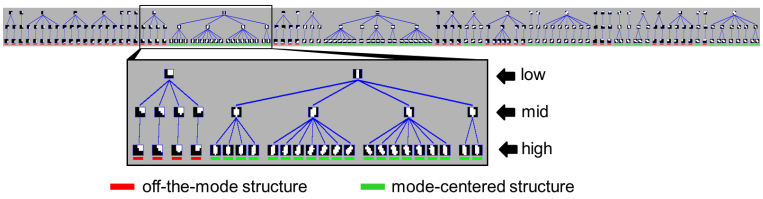

Figure 3. The hierarchically organized binary shape tree obtained for compact *cluster* delineation. Color labels indicate if the local structure represents a mode location.

we denote as *mode-centered* or *line-centered* samples, respectively. This state is stored in a flag c.

Figure 2 illustrates the sampling step and the generated representations. For each sampled location the window W is subdivided into a local grid. We define three spatial subdivision levels spanning a $n_i \times n_i$ grid, where $n_1 = 3$, $n_2 = 5$ and $n_3 = 7$ are used for a low-, mid- and high-resolution representation, respectively. At a given sampled location for each resolution we compute the rate of binary foreground pixels within each grid cell relative to the cell area. If the foreground pixel rate exceeds a preset threshold T ($T = 0.5$), we set the current cell label l_j to 1, otherwise 0, generating a binary label vector $\mathbf{l}_i$ for each resolution level i. Thus for each sampled location we represent the local structure by $S = \{\{\mathbf{l}_i\}_{i=1..3}, \mathbf{v}, c\}$, where $\mathbf{l}_i$ is the binary label vector at multiple resolutions, $\mathbf{v}$ is the polygonal contour information in a position-normalized (relative to the window center) and scale-normalized form, and c is the flag indicating whether the local sample is centered. For *lines* the stored set of attributes becomes $S = \{\{\mathbf{l}_i\}_{i=1..3}, \mathbf{t}, c\}$, where $\mathbf{t}$ contains the position- and scale-normalized end point coordinates for centered (c=1) structures. If end point coordinates are outside of the sampling window, they are still stored, but used only to determine the direction towards the end points in the later delineation step.

The obtained pool of binary signatures can be easily quantized by binning. Since low-resolution structures are simplified variants for many mid- and high-resolution shapes, therefore binning generates a tree-like hierarchy. The low-resolution layer represents prototypical shapes, while higher resolutions encode more detailed and specific structures. A typical tree obtained for the compact elliptic cluster representation is shown in Figure 3.

Shape delineation: The learned hierarchically organized codebook is used to efficiently find a best matching binary shape for a given image patch around a hypothetical mode location in an image I containing a two-dimensional distribution. Following steps are carried out:

Step 1: Three integral images are computed from I, $I \cdot x$ and $I \cdot y$ where x and y denote image coordinates for a given pixel. These integral images are used to compute area-sums and first-moments in order to carry out fast Mean Shift iterations from a set of detected local maxima, similarly to [1].

Step 2: At a potential mode location an analysis window

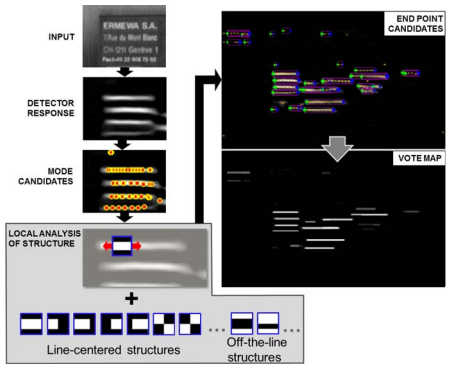

Figure 4. Our text line segmentation scheme. A text detector response map is analyzed locally by a hierarchically structured set of binary shapes. Best-fitting local structures, such as line segments and line extremities (top-right image) are used in a voting procedure to generate a text line probability map (bottom-right image).

of size W (manually set parameter) is defined. Using the same subdivision scheme, as in the learning step, the density measure D is computed for each resolution level i for the binary structure $\mathbf{l}_i$:

$$D_i(\boldsymbol{l}_i \mid I) = \frac{1}{A_F} \sum_{\{x,y \in C \mid \boldsymbol{l}=1\}} I(x,y) \; - \; \frac{1}{A_B} \sum_{\{x,y \in C \mid \boldsymbol{l}=0\}} I(x,y), \qquad (1)$$

where A_F and A_B denote the binary foreground/background areas within a cell C of the local grid, respectively. The above equation can be evaluated very efficiently by means of the same integral image which was used in Step 1 for area-sum computation.

Step 3: The best matching codebook entry at resolution i maximizes the density within the hypothesized foreground region, while minimizing the density in the hypothesized background region:

$$\boldsymbol{l}_i^* = \arg\max_{\boldsymbol{l}} \; D_i(\boldsymbol{l}_i \mid I). \qquad (2)$$

Evaluating the above measure at each resolution in the shape tree guides the selection towards the best matching high resolution candidate at the leaf-node of the tree. The stored polygonal contour information (after translation and scaling) can be used to delineate the underlying local distribution.

For *line* distributions (see Figure 4) the delineation requires additional steps. Finding the best matching structure at a mode location is complemented by an iterative line ending search scheme. If the local structure indicates a line-internal (i.e. between two end points) segment, the analysis window is translated to the left and to the right (using the stored line ending information $\mathbf{t}$) in a recursive manner until a line ending structure is found. Since the search is bi-directional, the scheme always generates a pair of left and right end point candidates (see Figure 4). Between the hypothesized end

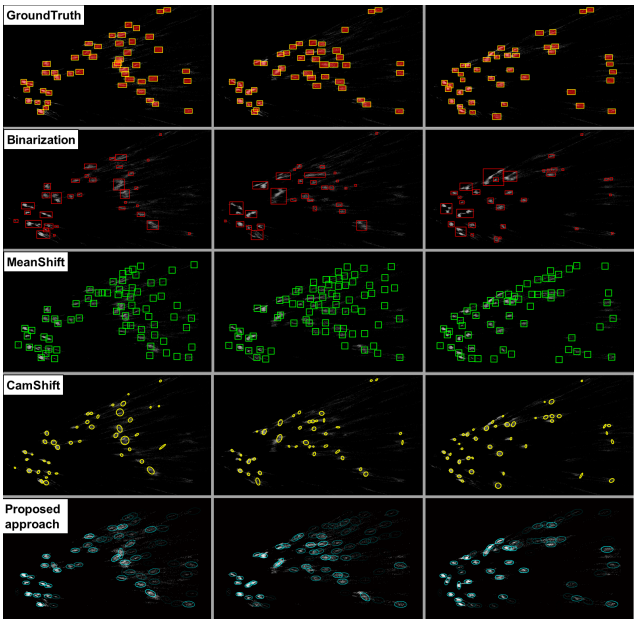

Figure 5. Annotated data (top row) and clustering results for four delineation schemes. Binarization (second row) shows obtained segmentation superimposed onto the original distribution. Our proposed approach is shown in the last row. Best viewed in color.

points we cast a vote in an accumulator image, thus multiple end point candidates generate a clearly defined and spatially well separated multi-line vote distribution which can be easily segmented afterward.

4. Results and discussion

We present delineation results for two applied cases to show the practical relevance of the proposed scheme. Cluster delineation is presented for human detection in occupancy maps (similarly to [7]), and line structure segmentation is demonstrated for a multi-line text segmentation task.

Human detection by occupancy map clustering: The occupancy maps are generated by passive stereo depth sensing, where depth data is projected orthogonal to the ground plane (see Figure 6). For experiments, we selected an image sequence depicting a crowded scene, where humans in the computed occupancy maps appear as noisy peaks embedded into clutter. Delineating individual clusters in the occupancy map, and reprojecting detection results into the original image result in images shown in Figure 6. We use the computed maximum density $D_3\left(\boldsymbol{l}_3^* \mid I\right)$ of Eq. 1 at the highest shape resolution as a confidence measure, thus the superimposed elliptic shapes exhibit a confidence-driven variable brightness.

In order to quantify the delineation quality, we manually annotated 165 frames of the crowd image sequence, in total representing 2762 annotated clusters linked to human individuals. Data clustering is performed by multiple techniques: binary thresholding followed by a morphological opening operation for suppressing noise, Mean Shift

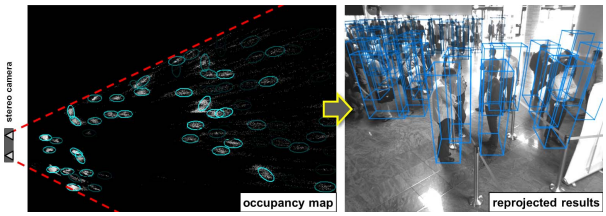

Figure 6. Figure demonstrating the task-specific relevance of generated clustering results. Left: the occupancy map (top-view of the scene) is shown with detected clusters. Right: clustering results shown as reprojected bounding boxes in the corresponding video frame of a crowded scene. Faces and signs are manually blurred to render them anonymous.

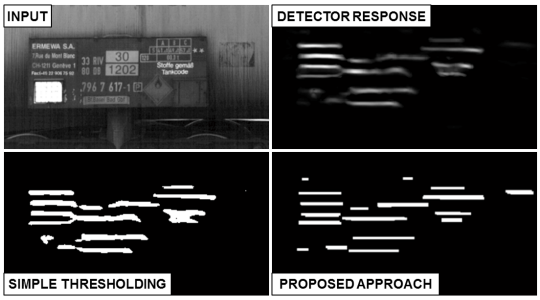

Figure 7. Comparison of text line segmentation results. The top row shows the image input and a computed text detector response map, respectively. The bottom row displays line segmentation results by simple thresholding and by our proposed scheme.

mode seeking [4] using a fixed window size, scale adaptive CamShift [1] and our proposed delineation scheme. The latter three methods all are based on mode seeking, thus they share the same sensitivity parameters for a fair comparison. The annotations and the obtained clustering results are shown in Figure 5.

Quantitative analysis was performed in terms of a Precision-Recall characterization, employing a bounding-box overlap criterion and a one-to-one match between the results of a given method and the annotations as reference data. Precision is referred to how many delineated clusters are relevant and equals to $\frac{tp}{tp+fp}$. Recall is referred to what fraction of the relevant clusters was found and equals to $\frac{tp}{tp+fn}$. tp, fp and fn are the number of true, false and missed cluster delineations, respectively. Table **1** displays the obtained quantitative results for all methods.

As it can be seen from the qualitative and quantitative results, binarization, unsurprisingly, does not cope well with multiple nearby structures and noise, and it undersegments nearby peaks and misses weakly defined structures. Mean Shift locates most true density maxima, but it exhibits no specificity towards cluster shapes thus it also locates mode candidates within the noise background. CamShift, due to its stability criterion, shows more robustness with respect to noise, but it often estimates the scale incorrectly: sometimes it converges too early yielding too small clusters, sometimes, when nearby peaks are present, it estimates too large clusters. Our proposed approach appears to generate the

Performance measure	Binarization	Mean Shift [4]	Cam Shift [1]	Proposed
Recall (R)	0.52	0.95	0.81	0.92
Precision (P)	0.86	0.76	0.89	0.87
F-measure (F)	0.65	0.84	0.85	**0.89**

Table 1. Obtained quantitative results for the cluster delineation task for the annotated set of occupancy maps (see Figure 5).

most accurate orientation-adaptive elliptic shapes, nevertheless, in presence of nearby structures, occasionally it fails to select the correct binary shape. A possible reason for this deficiency is that our local model captures shape representations for isolated elliptic shapes, but it does not include the joint distribution of multiple nearby shapes. Although joint shape models would span a vast space of configurations, the coarse-to-fine spatial discretization would permit to handle this combinatorial complexity. This aspect will be investigated by further work.

Text line segmentation: Figure 7 depicts our obtained results for text line segmentation. As it can be seen, simple thresholding of the text detector response results in under-segmented nearby text lines, and also suppresses weak signals in the response map. Our proposed scheme delineates all valid text lines correctly, and in addition, it also detects subtle patterns exhibiting a line-like structure. This latter property of the method can be beneficial if the analysis targets high recall and at the same time it must eliminate clutter and noise. The presented results are limited to horizontal line segments, but the scheme can be easily extended towards rotated line structures by employing corresponding training data.

Run-time performance: Our proposed approach runs at $22 fps$ when performing cluster delineation in occupancy maps (see Figure 6) with a resolution of 1246×728 pixels, using a modern PC and a partially optimized C++ code.

5. Conclusions

We present a fast two-dimensional cluster delineation scheme employing a coarse-to-fine shape codebook to enforce prior shape information upon analyzing a local distribution. We demonstrate for two applied examples the detection of relevant structures in heavily multi-modal and noisy distributions at a fast computational speed. The current implementation employs a fixed-scale analysis window, but the method inherently allows for scale variations. This aspect represents a future work towards a non-parametric analysis of local structures.

6. Acknowledgements

This work was supported by the Vision+ Project under the COMET program of the Austrian Research Promotion Agency (FFG); the European Unions 7^{th} Framework Programme for research, technological development and demonstration under grant agreement no.312583; and the research initiative Mobile Vision funded from the AIT Austrian Inst. of Technology and the Austrian Federal Ministry of Science, Research and Economy HRSM programme (BGBl.II Nr.292/2012).

References

[1] C. Beleznai, P. Sommer, and H. Bischof. Scale-adaptive clustering for object detection and counting. In *Proc. IEEE Workshop on Performance Evaluation in Surveillance and Tracking*, 2008. 2, 3, 4, 5

[2] G. R. Bradski. Computer vision face tracking for use in a perceptual user interface. *Intel Technology Journal*, (Q2):15, 1998. 1, 2

[3] J. Chua, I. Givoni, R. Adams, and B. Frey. Learning structural element patch models with hierarchical palettes. In *Conference on Computer Vision and Pattern Recognition*, pages 2416–2423, June 2012. 2

[4] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on PAMI*, 24:603–619, 2002. 1, 2, 4, 5

[5] P. Dollár and C. L. Zitnick. Structured forests for fast edge detection. In *Proc. IEEE International Conference on Computer Vision*, pages 1841–1848, 2013. 2

[6] D. Gavrila and V. Philomin. Real-time object detection for "smart" vehicles. In *Proc. IEEE Int'l Conf. on Computer Vision*, pages 87–93, 1999. 2

[7] O. H. Jafari, D. Mitzel, and B. Leibe. Real-time RGB-D based people detection and tracking for mobile robots and head-worn cameras. In *International Conference on Robotics and Automation*, pages 5636–5643, 2014. 2, 4

[8] Y. Jin and S. Geman. Context and hierarchy in a probabilistic image model. In *Conference on Computer Vision and Pattern Recognition*, pages 2145–2152, 2006. 2

[9] P. Kontschieder, S. R. Bul, H. Bischof, and M. Pelillo. Structured class-labels in random forests for semantic image labelling. In *ICCV*, pages 2190–2197. IEEE, 2011. 2

[10] B. Leibe, E. Seemann, and B. Schiele. Pedestrian detection in crowded scenes. In *Conference on Computer Vision and Pattern Recognition*, pages 878–885, 2005. 2

[11] A. Neubeck and L. Van Gool. Efficient non-maximum suppression. In *Proc. International Conference on Pattern Recognition - Volume 03*, pages 850–855, 2006. 2

[12] M. D. Rodriguez and M. Shah. Detecting and segmenting humans in crowded scenes. In *Proc. of the 15th International Conference on Multimedia*, pages 353–356, 2007. 2

[13] R. Rothe, M. Guillaumin, and L. Van Gool. Non-maximum suppression for object detection by passing messages between windows. In *Proc. Asian Conf. on Computer Vision*, Nov. 2014. 2

[14] C. Sun and P. Vallotton. Fast linear feature detection using multiple directional non-maximum suppression. In *18th International Conference on Pattern Recognition*, pages 288–291, 2006. 2