Unsupervised Object Discovery and Segmentation in Videos

Samuel Schulter[1]
schulter@icg.tugraz.at
Christian Leistner[2]
christian.leistner@microsoft.com
Peter M. Roth[1]
pmroth@icg.tugraz.at
Horst Bischof[1]
bischof@icg.tugraz.at

[1] Institute for Computer Graphics and Vision
Graz University of Technology
Graz, Austria

[2] Microsoft Photogrammetry
Graz, Austria

Introduction: Unsupervised object discovery (UOD) is the task of finding repeating patterns and common visual concepts across an unsorted set of images without any human supervision. These concepts should describe objects, like pedestrians or cars, and stuff, like road or sky. Once discovered, this information opens several interesting applications like summarization and filtering of visual content, discovering novel or unusual visual patterns, or reducing human annotation effort. As the amount of visual data grows exponentially and human annotation becomes costly, such applications are getting more and more important.
However, existing UOD approaches typically build on still images and have to rely on prior knowledge to yield accurate results. In this work, we propose a novel *video-based approach*, allowing also for exploiting motion information, which is a strong and physically valid indicator for foreground objects, thus, tremendously easing the task (see Figure 1).

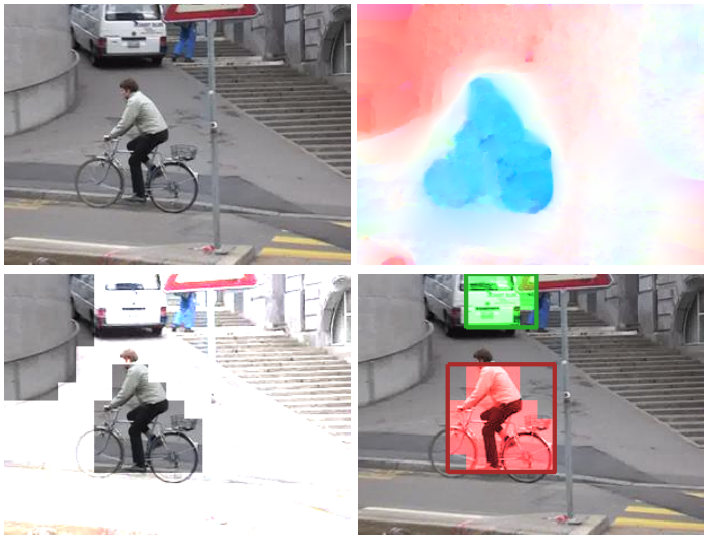

Figure 1: Given unlabeled video sequences *(top left)*, we first estimate the optical flow *(top right)* and calculate the motion segmentation *(bottom left)*. Then, using these segmentations we cluster the different object categories and learn appearance models. These models are finally used to discover objects in videos, where even static, non-moving objects as the car (green) can be found *(bottom right)*. Illustration for only one frame.

Approach: Our UOD approach consists of several building blocks as illustrated in Figure 2. The input is a set of videos and the number of categories k to be discovered, which is the only supervision necessary.
We start by calculating optical flow and perform a motion segmentation with our CRF formulation. This gives an initial set of tentative object proposals without any semantic information, yet. Our approach can then group those proposals into k semantically similar clusters and learn appearance-based object models for each cluster on two different abstraction levels (*local* and *holistic*) in a common framework. Given those object models, we can classify each frame of the input videos and calculate probability maps for each category. Given the additional semantic information from the learned classifiers, we formulate a joint CRF that allows for exploiting both, motion and semantic information. The CRF is now able (i) to confirm object proposals already identified via motion and to give them a semantic label, (ii) to remove outlier proposals from the first step, and (iii) also to discover new *static* objects having no motion information.

The output of our approach is thus a segmentation of each video frame into either background or one of k categories. Additionally, our approach also returns fully trained object detectors for each of the discovered categories, which can readily be applied to still images.

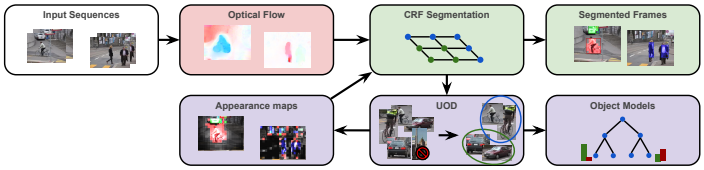

Figure 2: Building blocks of our Unsupervised Object Discovery approach from videos.

Experiments: To demonstrate the benefits of additionally using motion information for unsupervised object discovery, we run 3 different experiments: (i) Unsupervised object discovery in videos, where we show that our approach can successfully discover and discriminate between different categories. (ii) Video retrieval, where we demonstrate the generalization power of the learned appearance models. (iii) Object detection, where we show that our unsupervised learned models can even be applied to still images and give reasonable results compared to the fully supervised models.

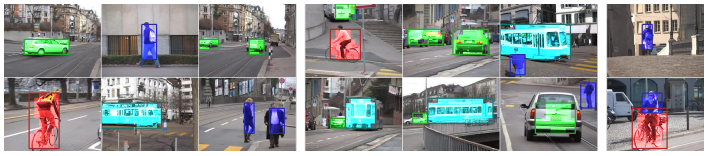

Figure 3: Some qualitative results of our Unsupervised object discovery approach.

In Figure 3, we show some qualitative results from the first experiment on UOD. Our approach segments the main objects in the sequences (first block) and is also able to identify static objects (second block). The last block illustrates two failure cases, which, however, also demonstrate the power of the part-based Hough Forests, as it actually separates the person and the bicycle (bottom row).

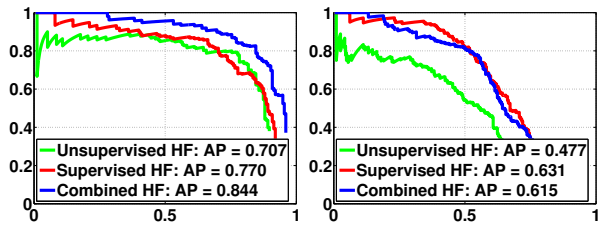

Figure 4: Precision-recall curves of all models on ETHZ-cars (left) and TUD-pedestrian (right).

In Figure 4, we compare our unsupervised trained models with the fully-supervised models and also a combination thereof on two different datasets. As can be seen from the average-precision curves, our unsupervised trained models yield reasonable results compared to their fully supervised competitors. On one data set, the combined approach can even outperform the supervised one.