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

Abstract

Unsupervised object discovery is the task of finding recurring objects over an unsorted set of images without any human supervision, which becomes more and more important as the amount of visual data grows exponentially. Existing approaches typically build on still images and rely on different prior knowledge to yield accurate results. In contrast, 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. In particular, we show how to integrate motion information in parallel with appearance cues into a common conditional random field formulation to automatically discover object categories from videos. In the experiments, we show that our system can successfully extract, group, and segment most foreground objects and is also able to discover stationary objects in the given videos. Furthermore, we demonstrate that the unsupervised learned appearance models also yield reasonable results for object detection on still images.

1 Introduction

The ever-growing amount of images and videos asks for automatic approaches that analyze and summarize the upcoming data, as human annotation becomes too costly or even infeasible. Unsupervised object discovery (UOD) systems tackle this problem by finding common visual concepts across an unlabeled set of images. These concepts should then describe *objects*, like cars or pedestrians, and *stuff*, like road or sky. Once discovered, this information opens several interesting applications such as (i) reducing human labeling efforts and costs when training classifiers, (ii) avoiding user-specific bias in annotation tasks, (iii) discovering novel or unusual visual patterns, and (iv) summarizing and filtering visual content.

Previous approaches tackled the task of UOD mainly using collections of still images [13, 20, 27], which are either based on topic modeling or clustering methods. Even though showing promising results, it is still hard to separate *object*- from *stuff*-regions without any prior knowledge. Moreover, *object* regions describe the most interesting visual concepts for many

tasks, *e.g.*, object retrieval [22, 29]. Thus, recent works [19, 20] are mainly focused on *objects* by exploiting prior knowledge like object detectors or saliency measures.

In this work, we show that the use of video data can tremendously ease UOD. Compared to still images, videos offer several advantages: (i) It is relatively easy to segment an object from the background using motion and observed temporal consistency, (ii) videos capture a higher variability of an object's appearance, and (iii) a small set of videos already provides a large number of training images. Moreover, there are rich and easily accessible sources for realistic video material such as *YouTube* or *Vimeo*, and it has been reported in literature [3, 25, 31] that learning from continuous image sequences is also biologically plausible. Humans tend to learn and categorize moving objects first, *i.e.*, foreground vs. background, learn valid transformations, *etc.*, and then gradually extract general knowledge that also allows for improving the recognition skills on stationary objects.

Therefore, we propose an unsupervised method for object discovery in videos. Not assuming any prior knowledge about location, number or category of objects, we take unlabeled videos as input and estimate a motion segmentation. We use a Conditional Random Field (CRF) [17] formulation with potentials based on estimated optical flow fields. Given the motion segmentation, we extract tentative object proposals and discover semantically different object classes via a robust clustering approach. Based on a recently popular object detection approach, Hough Forests (HF) [12], we simultaneously learn appearance models on two different abstraction levels, one for segmentation (pixels) and another one for object detection (bounding boxes). We integrate all information cues (motion and two appearance cues) into a combined CRF formulation to discover objects in videos, even static ones. Finally, the output of our approach is two-fold: (i) The videos segmented into semantic labels and (ii) a (fully unsupervised) trained object detector instantly applicable to *still images*.

In our experiments, we show the benefits of additionally using motion information for UOD. We demonstrate that our approach can successfully identify and cluster objects in videos and further show the generalization power of the unsupervised learned object models for video retrieval and even for detection on still images.

2 Related Work

Unsupervised object discovery was mainly studied in context of still images. Existing works are either based on latent topic models [27, 30] or clustering algorithms [13, 19]. Sivic *et al.* [30] proposed to use Probabilistic Latent Semantic Analysis (PLSA) and Latent Dirichlet Allocation to separate images of different object categories. Russell *et al.* [27] extended this approach by using multiple image segments instead of the original images as the equivalent of documents in the topic model. As a drawback, all these methods share the same concept of model assumption. However, image categories are arranged in complex and unknown shapes, making designing explicit models difficult.

An alternative research direction, which is more versatile in handling structured data, builds on similarity-based methods. Frey and Dueck [9] applied their affinity propagation algorithm [11] for unsupervised image categorization. Grauman and Darrell [13] developed partially matching image features to compute image similarity and used spectral clustering for unsupervised category learning. As the semantic level grows, measuring similarities between images becomes the main difficulty. Lee and Grauman [20] apply curriculum learning [4] in order to discover visual categories. Similar to [19], [20] assumes some prior knowledge about previously learned categories, background and objectness [1], and iteratively tries to discover object clusters by concentrating on "easy" samples first.

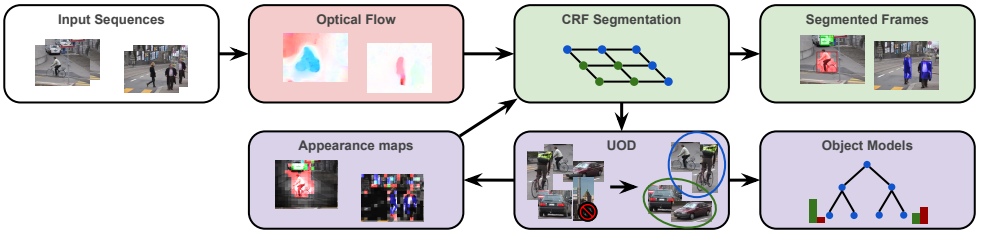

Figure 1: Overview: Given a set of videos, we first calculate optical flow and a CRF-based motion segmentation, returning a set of object proposals. Next, we remove outliers, cluster k categories, and train appearance classifiers for each cluster (UOD Block). The output of the learned classifiers on the videos is integrated into the CRF formulation to finally discover and label the objects in the videos (even static ones).

In contrast, studying *videos* for UOD was only of limited interest up to now. For instance, Liu and Chen [21] proposed a latent topic model for UOD in videos. However, the method concentrates only on relatively small objects in low-resolution images, a typical setup for surveillance scenarios. Later the same authors proposed a video retrieval system that estimates similarities between videos based on local information of the object of interest [22].

Most recently, also weakly-supervised learning from videos has been of interest. Prest *et al*. [26] proposed to jointly extract moving objects out of a set of *equally labeled* videos to learn appearance-based models for object detection on still images. Similarly, Hartmann *et al*. [14] perform spatio-temporal semantic segmentation of objects from videos with (noisy) labels. Although both approaches demonstrate promising results, the videos still have to be labeled and object categories cannot be *discovered*.

Our approach can also be motivated by findings from cognitive sciences [3, 31] on how humans learn to recognize objects. Long-term studies on previously blind humans who gained their visual senses after successful surgeries clearly revealed (i) that visual recognition is learned, (ii) that observation of motion and transformation of foreground objects are important, and (iii) that labels are less important than previously expected, (*cf*. [25]).

3 Unsupervised Object Discovery in Videos

In the following, we describe our approach to unsupervised object discovery from videos, which is illustrated in Figure 1. The input is an unordered set of n videos $\mathcal{V} = \{V_i\}_{i=1}^{n}$, each consisting of m_i frames $V_i = \{f_i^1, \ldots, f_i^{m_i}\}$, and the only supervision is given by the number of categories k to be discovered. The ultimate goal is to localize all objects (moving *and* static) and to assign a semantic label to each of them. The output 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.

We start by estimating motion information via optical flow (Sec. 3.1) and perform a motion segmentation with our CRF formulation (Sec. 3.2). 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 sets and learn appearance-based object models for all sets on two different abstraction levels (*local* and *holistic*) in a common framework (Sec. 3.3). Given those object models, we can classify each frame of the input videos and calculate probability maps for each category.

Next, our proposed joint CRF formulation allows for exploiting both, motion and semantic information, and generates more elaborate object proposals. With the additional 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.

3.1 Optical Flow and Pre-processing

As a first step, we calculate an optical flow field $\mathcal{F}_i^j$ for each frame f_i^j in all video sequences V_i (see Fig. 1). We use the method of [33], which is quite robust and efficient, even using a standard consumer *GPU*. We get a dense optical flow field $\mathcal{F}_i^j = \{\mathbf{v}_k\}_{k=1}^K$, where $\mathbf{v}_k$ is the flow vector at pixel k in frame f_i^j and K is the number of pixels in that frame.

To make the motion estimation more robust, we have to compensate for camera motion. This is important as many video sequences are captured from non-stationary cameras, thus, non-moving background regions could yield higher responses than moving objects. For that purpose, we take a small part of the optical flow's border region to robustly estimate an affine model of the camera movement via RANSAC. We observed that, for our task, this gives similar results as well-known motion estimation approaches (*e.g.*, [7]), but being simpler and much faster.

As we have to deal with huge amounts of data, we reduce the computational complexity of the CRF-based segmentation task by discretizing the image in a coarser pixel-grid with a fixed size. Please note that we can also calculate boundary-aligned super-pixels, but as we do not focus on pixel-accurate segmentations, we stick with the more efficient pixel-grid that can also be seen as regular super-pixels. Throughout the paper, we refer to the pixel-grid as a set of super-pixels $\mathcal{S}_i^j = \{s_l\}_{l=1}^L$ for each frame f_i^j, where L is the number of super-pixels in a single frame.

3.2 Object Extraction and Segmentation via CRF

This section describes our CRF formulation that serves two purposes: (i) *Motion segmentation* if no semantic information from appearance models is given and (ii) *semantic segmentation* as soon as all information cues of our approach are available, *i.e.*, after having discovered object categories (see Sec. 3.3). We define the input as the set of videos V_i, the computed optical flow fields $\mathcal{F}_i^j$, the super-pixels $\mathcal{S}_i^j$, and the appearance information of all frames f_i^j. Recap that appearance information is not available in the first iteration. The output is a semantic segmentation of the given video frames.

To solve the CRF, we construct a graph $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$ in each input video frame f_i^j. The vertices $\mathcal{V}$ correspond to the super-pixels s_l in a frame f_i^j and take a label $l \in \mathcal{L}$, where $\mathcal{L}$ consists of $k+1$ labels (k categories and a *background* label). The edges $\mathcal{E}$ correspond to neighboring super-pixels, where we define two superpixels as neighbors if they share a common boundary. We also link two superpixels not only spatially but also temporally, *i.e.*, across two frames f_i^{j-1} and f_i^j, in order to account for the space-time coherence. The CRF finds the optimal labeling x of all super-pixels s_l by minimizing the energy

$$E(x) = \sum_{s_l \in \mathcal{V}} \Phi(s_l) + \sum_{(s_l, s_k) \in \mathcal{E}} \Psi(s_l, s_k) \,, \tag{1}$$

where $\Phi(s_l)$ is the unary potential for super-pixel s_l, and $\Psi(s_l, s_k)$ is the pair-wise potential between super-pixels s_l and s_k.

Unary potentials: We formulate the unary potentials such that they integrate both, motion and appearance information. We thus define them as the sum over different cues:

$$\Phi(s_l) = \sum_c \gamma_c^U \cdot \Phi_c(s_l) \,, \tag{2}$$

where c is either motion (m), super-pixel appearance (spa), or object appearance (oa). The latter two cues are the two different abstraction levels of appearance information given by our classifiers (see Sec. 3.3); γ_c^U are positive steering factors that sum up to 1.

We define $\Phi_m(s_l)$ to be dependent on the norm of the median over all flow vectors $\mathbf{v}$ within the superpixel s_l. Higher norms indicate higher probabilities for objects (*i.e.*, one of the k categories) and lower values indicate higher probabilities for background. Note that we spread the foreground probabilities equally to all k categories to fill the label space, which allows us for using exactly the same CRF solver for both scenarios, *motion segmentation* and *semantic segmentation*. We add a constant probability η for foreground objects to avoid having zero probability for static objects, *i.e.*, superpixels with a zero norm median flow vector. The costs for the motion cue can thus be defined as

$$\Phi_m(s_l) = -\log\left(\eta + \frac{\text{med}(\|\mathbf{v}(s_l)\|)}{\max_l \text{med}(\|\mathbf{v}(s_l)\|)}\right) \,, \tag{3}$$

where $\text{med}(\cdot)$ denotes the median operator. Note that this is the only potential active in the first iteration (*i.e.*, *motion segmentation*).

To define the appearance-based costs $\Phi_{spa}(\cdot)$ and $\Phi_{oa}(\cdot)$, we assume that we are given the probability maps $p_{spa}(\cdot)$ and $p_{oa}(\cdot)$ in the range $[0,1]$ from the appearance-based classifiers (Sec. 3.3) for all frames f_i^j. Then, we can define $\Phi_{spa}(s_l) = -\log(p_{spa}(s_l))$ and $\Phi_{oa}(s_l) = -\log(p_{oa}(s_l))$ (see Sec. 3.3) to be dependent on the output of the super-pixel level and the object level classifiers, respectively.

Pair-wise potentials: We use contrast-sensitive pair-wise potentials for penalizing label transitions based on color *and* motion differences between neighboring superpixels s_l and s_k. We compute the mean *RGB* color vector for both superpixels and use the normalized difference to calculate a weight w_{kl}^{color} between s_l and s_k. Motion weights w_{kl}^{motion} are calculated in the same way based on the mean flow vectors of each superpixel. As for the unary terms, we also compute a weighted sum between the two pair-wise factors $\sum_c \gamma_c^P \cdot w_{kl}^c$, where c is either *color* or *motion*. Throughout all our experiments, we equally weight the influence of both factors.

3.3 Unsupervised Learning of Appearance Models

Next, we discuss the unsupervised learning of object models from video data, which includes grouping of object proposals into semantically similar clusters and learning appearance-based models to further improve the overall discovery process. As can be seen from Figure 1, the only input to our unsupervised discovery approach are the object proposals from the initial motion segmentation, *i.e.*, some proposals that might stem from moving foreground objects. The goal in this step is (i) to remove outlier proposals, (ii) to identify k semantically meaningful clusters, (iii) to learn appearance-based models for each of the discovered clusters, and (iv) to apply those models to all input frames.

Outlier Removal: To remove outliers, we exploit the potential of video data (compared to single image data), *i.e.*, the temporal coherence of moving objects. We thus assume that each object is moving smoothly through space and time. This allows us for easily detecting

frames where motion segmentation fails and to remove them. Of course videos often consist of different shots, thus breaking the smoothness assumption. However, by using a simple shot boundary detection as in [26] the videos can be split and the problem is alleviated.

We define the object proposals given by motion segmentation as P_m^{ij}, $m = 1,\ldots,M$, where ij associates the proposal with frame f_i^j and M is the number of extracted proposals. Each proposal is described with a bounding box, found by calculating the tightest bounding box around the region that was marked "moving" during motion segmentation. Frames f_i^j that do not contain any proposals are discarded immediately.

To effectively remove outliers, we first gather statistics of all proposals P_m^{ij}. We collect the x- and y-coordinates of the center and the width w and height h of the bounding boxes over time, *i.e.*, the temporal evolution of the proposals. As we assume smooth motion of the object, the gathered statistics of the proposals should also vary smoothly. Thus, we solve a simple line-fitting problem via RANSAC for each of the collected statistics over time and discard all frames f_i^j that are identified as outlier. The video shots are typically short enough such that a simple linear line fit suffices to model the trajectory of the statistics.

Clustering: After having removed outliers, we are left with a set of $\hat{M}$ object proposals $P_{\hat{m}}^{ij}$, $\hat{m} = 1,\ldots,\hat{M}$, which should be clustered into k categories. For that purpose, we describe each proposal with a strong appearance feature, namely a Bag-Of-Words (BoW) model built on dense SIFT features using a 300-dimensional codebook and a standard spatial pyramid [18]. Based on the $\mathcal{X}^2$-distance, we calculate a similarity matrix S of all proposals $P_{\hat{m}}^{ij}$ to employ a spectral clustering approach [8], assigning each proposal to one of the k clusters.

Appearance models: The final step of our unsupervised object discovery process is to learn appearance models for each cluster to provide the CRF with semantic class probabilities in the next iteration. Most standard semantic segmentation approaches only use a classifier on the (super-) pixel level for providing unary potentials (*e.g.*, [28]). However, there is a strong trend to use higher-level information like object detector outputs [16]. We also observed that this is useful in our case.

In fact, we employ the Hough Forests (HFs) [12] to capture both levels. HFs model an object as a set of 16×16 image patches having an offset vector pointing to the object center. Random Forests [2] are trained with adapted splitting criteria for minimizing both, class uncertainty and offset variance of patches. During inference, test patches traverse down the trees to leaf nodes that store offset vectors from training. These offset vectors then vote for tentative object centers in a Hough space. For more details we refer the reader to [12].

We extract features densely on all frames f_i^j as described in [12]. For the super-pixel level, we crop a feature patch on the region with double the size for each superpixel s_l (thus capturing some context information) and resize it to 16×16 pixel. Here, we set the offset vector d to a zero vector. For the holistic level, we also follow [12] and resize each object proposal to a common height (100 pixel) and randomly extract 16×16 feature patches, including an offset vector d pointing to the object center. To gather negative samples, we collect features on background regions according to the motion segmentation. We thus can use the same feature representation for both levels. Due to the different scales and the fact that we cannot exploit the offset vector regression for the superpixel level (we only use the class uncertainty criterion [12]), we train two forests for the two different levels.

In order to integrate this new information into the final CRF segmentation, we apply both classifiers on all frames f_i^j. For the superpixel level, we extract features for all super-pixels s_l as in the training phase and apply the learned classifier to get class estimates $p_{spa}(s_l)$. For the holistic level, we densely evaluate the learned HF on the frames f_i^j and use the resulting

Hough maps to compute the class estimates $p_{oa}(s_l)$ for each super-pixel s_l. For that purpose, we compute the mean confidence $\bar{\omega}_{s_l}$ within the area of s_l and transform the confidence into a probabilistic output as $p_{oa}(s_l) = \frac{1}{1+exp(-\bar{\omega}_{s_l})}$. Although the voting maps only show peaks at tentative object centers, the semantic information is propagated to object boundaries by the subsequent CRF with the help of superpixel, color, and flow information. The holistic classifier thus rather indicates semantic information for segments. Uur experimental evaluation shows that this information helps to improve the overall performance. Furthermore, these holistic models can be immediately used for prediction in still images, which we also evaluate in the next section.

4 Experimental Evaluation

To demonstrate the benefits of our approach, we run 3 different experiments: (i) Unsupervised object discovery – to demonstrate the ability of discovering and discriminating between categories; (ii) video retrieval – showing the generalization power of the learned appearance models; (iii) object detection on still images using the unsupervised learned models.

We validate our approach in the first two experiments with the data set presented in [23][1], which captures 4 object categories (*bicycle*, *car*, *pedestrian*, *streetcar*) in 96 videos and more than 7000 frames. Each video shows one prominently moving object, but also other objects in the background, like parking cars or pedestrians. The videos are recorded with a non-static hand-held camera, thus making the data set quite challenging.

Implementation Details: We use the features from the standard Hough Forest implementation [12] for both the superpixel- and the object-level, as described in Sec. 3.3. We train each Hough Forest with 10 trees for both levels, respectively, and fix the maximum depth to 15.

To solve the *CRF* formulation, we use the software from [5, 6, 15]. Unary and pair-wise potentials are defined as described in Sec. 3.2. We set the steering factors of the unary potentials in the first iteration to $\gamma^U = [1\ 0\ 0]$. In the second iteration, we increase the importance of the appearance cues and thus set $\gamma^U = [0.25\ 0.40\ 0.35]$. We tuned these steering parameters by hand and observed that they are quite insensitive to the final performance as long as each cue is given reasonable importance (≥ 0.1).

4.1 Unsupervised Object Discovery in Videos

First, we apply the proposed approach to the task of unsupervised object discovery in videos. The goal is to show that we can successfully discover different moving objects captured in the sequences, requiring only the number of categories, *i.e.*, $k = 4$. Furthermore, we want to demonstrate that our approach is also able to discover static objects in the background, *e.g.*, parking cars or bicycles.

We use *purity* [32] as quantitative performance measure, which is the percentage of correctly classified frames f_i^j, when each discovered cluster gets labeled with the majority class label of its assigned discovered frames. A single frame is correctly classified, if its largest discovered segment is assigned the correct label.

Although there is only limited related work also exploiting motion cues for UOD, we still compare with a standard UOD approach [27]. Since [27] failed without assistance of motion-cues, we additionally provide the system with our motion segmentations. This makes it similar to [21] and allows [27] for discovering k topics.

[1] We thank the authors for providing the data set.

Model	Purity
Ours (full)	75.1
Ours (sp. only)	72.3
Ours (holistic only)	69.4
Ours (no outlier rem.)	62.2
[21] $k=4$	52.0
[21] $k=5$	55.0

-	c1	c2	c3	c4
c1	**65**	05	12	06
c2	06	**88**	02	06
c3	13	06.1	**80**	04
c4	13	00.1	04	**84**

Model	Frame	Video
Ours (full)	65.9	73.9
[23]	74.3	87.4
[24] Appear	53.0	58.9
[24] Shape	74,4	88.4
[24] Comb.	81.4	94.5

Table 1: (left) Results of the UOD task (as *purity*); (middle) confusion matrix for *Bicycle* (c1), *car* (c2), *pedestrian* (c3), and *streetcar* (c4); (right) results of the retrieval task. [23] and [24] use weak-supervision; our approach just needs the number of categories k.

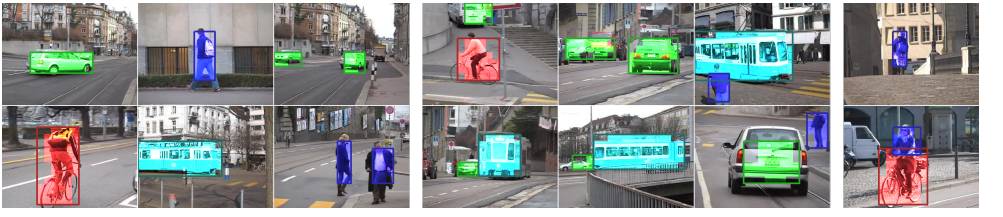

Figure 2: Qualitative results of our unsupervised discovery approach. The first block shows successful discoveries of moving objects, while the second block also shows discoveries of static objects in the background. The last block illustrates two failure cases.

Our quantitative results are summarized in Table 1. As can be seen, our approach yields quite good results, correctly classifying around 75% of all frames, and also outperforms [21] significantly. We further evaluate the influence of the different information cues and the proposed outlier removal process. Omitting the outlier removal gives poor results (around 62%), which can be explained by the fact that the categories *bicycle* and *person* do not form well separable clusters. We also provide the results when only one appearance cue, *i.e.*, either superpixel or holistic, is used. As can be seen, the superpixel information seems to be more informative, however, combining both cues gives the best score.

Table 1 also depicts the resulting confusion matrix. As expected, the categories *bicycle* and *person* show most confusion. This can be easily explained by the fact that persons always ride a bicycle in all videos of this data set. However, we still can classify 80% of the *bicycle* videos correctly, showing the discriminative power of our models.

We also show qualitative results in Figure 2. 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).

However, the scalability of this unsupervised approach with respect to the number of categories k has to be further evaluated and is left for future work.

4.2 Unsupervised Object Retrieval

Next, we apply our approach on the object retrieval task presented in [24]. The task is to learn the objects from the training videos and to assign each testing frame the correctly retrieved category. Following [24], we split the data in 72 training videos (24 per category) and 24 testing videos (8 per category). We average our results over 3 independent runs. As in the previous experiment, we assign each frame the label of the largest retrieved segment. Please note that the test sequences are independent and have never been seen during training.

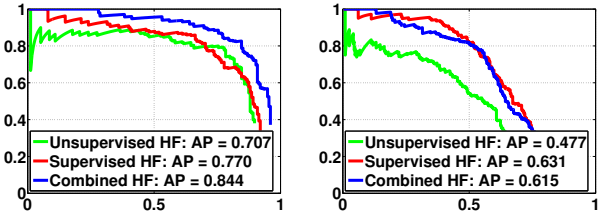

Figure 3: Precision-recall curves on ETHZ-cars (left) and TUD-pedestrian (right).

We measure the performance of all approaches by the retrieval rates per frame and per video, *i.e.*, the percentage of correctly classified frames and videos. As our approach is unsupervised and thus only finds k clusters, we assign each cluster the most frequently occurring label within its assigned frames to calculate a score. We compare our unsupervised approach with the weakly-supervised works [23, 24]. Both works present similar frameworks that segment moving objects based on tracking compositions of interest points, combined with appearance and shape information. Results are given for several variants of different feature types and combinations thereof. In contrast to our approach, these works need more supervision (*i.e.*, the label for each video) and additionally exploit shape features.

Table 1 depicts all results. The best result in [24] is obtained with a combination of shape and appearance features, where the shape feature is more important. Even though we do not use shape features and have less supervision, our approach yields reasonable results. Interestingly, our unsupervised ("appearance only") approach attains 13% better retrieval rates than the weakly-supervised ("appearance only") variant from [24]. However, our approach loses 15% compared to the best result of [24], but at a much lower level of supervision and without any shape information.

4.3 Recognition in Still Images

Finally, we show that our unsupervised learned (holistic) appearance models can also be applied to still images. The goal is to demonstrate that the unsupervised trained models yield comparable results to the original fully-supervised Hough Forest [12]. We further combine the different training sets and evaluate the combined HF.

We use the *TUD-pedestrian* data set and the *ETHZ-cars* data set for the evaluation of object detection. Our unsupervised approach (*Unsupervised HF*) is solely trained on the object proposals P_m^{ij} from the corresponding clusters (see Sec. 3.3); the supervised model (*Supervised HF*) on the fully-annotated training data sets (at bounding-box level). To train the combined Hough Forest (*Combined HF*), we used both data sets. For a fair comparison, all models are trained with the same settings for the forest: we use 10 trees, a maximum depth of 15, and 1000 random splitting functions. Please note that our unsupervised approach faces two main challenges: (i) Domain adaption, as the training and testing data is completely different, and (ii) little supervision, compared to the fully-supervised models.

Figure 3 depicts the results as precision-recall curves and average precision (AP) [10]. Our unsupervised models yield reasonable results on both data sets compared to the fully-supervised approaches, confirming the quality of our object discovery. For *TUD-pedestrian*, the combined model is slightly worse than the supervised model, as the test images mainly show pedestrians from a side-view unlike the additional unsupervised data collected from arbitrary views. However, for *ETHZ-cars*, the combined model can even outperform the supervised model, which is a motivating result as the unlabeled data comes for free.

5 Conclusions

We addressed the task of unsupervised object discovery in videos and showed that exploiting motion cues helps to drastically ease the task. We formulated an iterative process that exploits both, motion and appearance cues, via a joint CRF formulation to extract and segment objects. Our experiments show that the proposed method allows for discovering and grouping objects in unlabeled video sequences. Moreover, the unsupervised learned appearance models generalize well on unseen videos, can identify static objects having no motion, and can even be applied on still images.

Acknowledgement: The work was supported by the FFG projects Human Factors Technologies and Services (2371236) and Mobile Traffic Checker (8258408).

References

[1] Bogdan Alexe, Thomas Deselaers, and Vittorio Ferrari. What is an Object? In *CVPR*, 2010.

[2] Yali Amit and Donald Geman. Shape Quantization and Recognition with Randomized Trees. 9(7):1545–1588, 1997.

[3] Benjamin Balas and Pawan Sinha. Observing Object Motion Induces Increased Generalization and Sensitivity. *Perception*, 37(8):1160–1174, 2008.

[4] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum Learning. In *ICML*, 2009.

[5] Yuri Boykov and Vladimir Kolmogorov. An Experimental Comparison of Min-Cut/Max-Flow Algorithms for Energy Minimization in Vision. *PAMI*, 26(9):1124–1137, 2004.

[6] Yuri Boykov, Olga Veksler, and Ramin Zabih. Fast Approximate Energy Minimization via Graph Cuts. *PAMI*, 23(11):1222–1239, 2001.

[7] Thomas Brox and Jitendra Malik. Object Segmentation by Long Term Analysis of Point Trajectories. In *ECCV*, 2010.

[8] Wen-Yen Chen, Yangqiu Song, Hongjie Bai, Chih-Jen Lin, and Edward Y. Chang. Parallel Spectral Clustering in Distributed Systems. *PAMI*, 33(3):568–586, 2010.

[9] Delbert Dueck and Brendan J. Frey. Non-Metric Affinity Propagation for Unsupervised Image Categorization. In *ICCV*, 2007.

[10] Mark Everingham, Luc van Gool, Chris K. I. Williams, John Winn, and Andrew Zisserman. The Pascal Visual Object Classes (VOC) Challenge. *IJCV*, 88(2):303–338, 2010.

[11] Brendan J. Frey and Delbert Dueck. Clustering by Passing Messages Between Data Points. *Science*, 315(5814):972–976, 2007.

[12] Jürgen Gall and Victor Lempitsky. Class-Specific Hough Forests for Object Detection. In *CVPR*, 2009.

[13] Kristen Grauman and Trevor Darrell. Unsupervised Learning of Categories from Sets of Partially Matching Image Features. In *CVPR*, 2006.

[14] Glenn Hartmann, Matthias Grundmann, Judy Hoffman, David Tsai, Vivek Kwatra, Omid Madani, Sudheendra Vijayanarasimhan, Irfan Essa, James Rehg, and Rahul Sukthankar. Weakly Supervised Learning of Object segmentations from Web-Scale Video. In *ECCV-WS*, 2012.

[15] Vladimir Kolmogorov and Ramin Zabih. What Energy Functions Can Be Minimized via Graph Cuts? *PAMI*, 26:147–159, 2004.

[16] Lubor Ladicky, Paul Sturgess, Karteek Alahari, Chris Russell, and Philip H.S. Torr. What,Where & How Many? Combining Object Detectors and CRFs. In *ECCV*, 2010.

[17] John Lafferty, Andrew McCallum, and Fernando Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *ICML*, 2001.

[18] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In *CVPR*, 2006.

[19] Yong Jae Lee and Kristen Grauman. Object-Graphs for Context-Aware Category Discovery. In *CVPR*, 2010.

[20] Yong Jae Lee and Kristen Grauman. Learning the Easy Things First: Self-Paced Visual Category Discovery. In *CVPR*, 2011.

[21] David Liu and Tsuhan Chen. A Topic-Motion Model for Unsupervised Video Object Discovery. In *CVPR*, 2007.

[22] David Liu and Tsuhan Chen. Video Retrieval Based on Object Discovery. *CVIU*, 113 (3):397–404, 2009.

[23] Björn Ommer and Joachim M. Buhmann. Compositional object recognition, segmentation, and tracking in video. In *EMMCVPR*, 2007.

[24] Björn Ommer, Theodor Mader, and Joachim M. Buhmann. Seeing the Objects Behind the Dots: Recognition in Videos from a Moving Camera. *IJCV*, 83(1):57–71, 2009.

[25] Yuri Ostrovsky, Ethan Meyers, Suma Ganesh, Umang Mathur, and Pawan Sinha. Visual Parsing After Recovery From Blindness. *Psychological Science*, 20(12):1484–1491, 2009.

[26] Alessandro Prest, Christian Leistner, Javier Civera, Cordelia Schmid, and Vittorio Ferrari. Learning Object Class Detectors from Weakly Annotated Video. In *CVPR*, 2012.

[27] Bryan C. Russell, Alexei A. Efros, Josef Sivic, William T. Freeman, and Andrew Zisserman. Using multiple segmentations to discover objects and their extent in image collections. In *CVPR*, 2006.

[28] Jamie Shotton, Matthew Johnson, and Roberto Cipolla. Semantic texton forests for image categorization and segmentation. In *CVPR*, 2008.

[29] Josef Sivic and Andrew Zisserman. Video Google: A Text Retrieval Approach to Object Matching in Videos. In *ICCV*, 2003.

[30] Josef Sivic, Bryan C. Russell, Alexei. A. Efros, Andrew Zisserman, and William T. Freeman. Discovering Objects and Their Location in Images. In *ICCV*, 2005.

[31] Elizabeth S. Spelke. Principles of Object Perception. *Cognitive Science*, 14:29–56, 1990.

[32] Tinne Tuytelaars, Christoph H. Lampert, Matthew B. Blaschko, and Wray Buntine. Unsupervised Object Discovery: A Comparison. *IJCV*, 88(2):284–302, 2010.

[33] Manuel Werlberger, Werner Trobin, Thomas Pock, Andreas Wedel, Daniel Cremers, and Horst Bischof. Anisotropic Huber-L1 Optical Flow. In *BMVC*, 2009.