Multi-Class Video Co-Segmentation with a Generative Multi-Video Model

Wei-Chen Chiu Mario Fritz
Max Planck Institute for Informatics, Saarbrücken, Germany
{walon,mfritz}@mpi-inf.mpg.de

Abstract

Video data provides a rich source of information that is available to us today in large quantities e.g. from online resources. Tasks like segmentation benefit greatly from the analysis of spatio-temporal motion patterns in videos and recent advances in video segmentation has shown great progress in exploiting these addition cues. However, observing a single video is often not enough to predict meaningful segmentations and inference across videos becomes necessary in order to predict segmentations that are consistent with objects classes. Therefore the task of video co-segmentation is being proposed, that aims at inferring segmentation from multiple videos. But current approaches are limited to only considering binary foreground/background segmentation and multiple videos of the same object. This is a clear mismatch to the challenges that we are facing with videos from online resources or consumer videos.

We propose to study multi-class video co-segmentation where the number of object classes is unknown as well as the number of instances in each frame and video. We achieve this by formulating a non-parametric bayesian model across videos sequences that is based on a new videos segmentation prior as well as a global appearance model that links segments of the same class. We present the first multi-class video co-segmentation evaluation. We show that our method is applicable to real video data from online resources and outperforms state-of-the-art video segmentation and image co-segmentation baselines.

1. Introduction

Video data is one of the fastest growing resource of publicly available data on the web. Leveraging such resources for learning and making it accessible and searchable in an easy way is a big opportunity – but equally a big challenge. In order to leverage such data sources, algorithm must be able to deal with the unstructured nature of such videos which is beyond today's state-of-the-art.

Video segmentation has recently made great progress in improving on traditional segmentation algorithms. Motion and spatio-temporal structure in videos provide rich cues about potential object boundaries and independently moving objects. However, this approach has inherent limitations. As a single video might only expose a partial view, accidental similarities in appearance and motion patterns might lead to an ambiguous or even misleading analysis. In addition, performing video segmentation independently on each video of a video collection does not reveal any object class structure between the segments that would lead to a much richer representation.

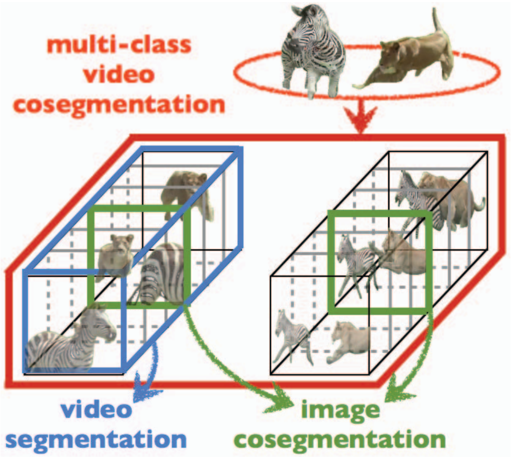

Figure 1. Our proposed multi-class video co-segmentation model addresses segmentation of multiple object classes across multiple videos. The segments are linked within and across videos via the global object classes.

We draw two conclusions. First, segmentations should be treated in a probabilistic framework in order to account for uncertainty. Second, a richer problem set should be investigated where the approach is enabled to reason across multiple video sequences in order to collect additional evidence that is able to link segments across videos.

Recently, two initial attempts [16, 5] have been made to approach such a video co-segmentation task. But these approaches make quite strong assumptions. A binary foreground vs. background segmentation is assumed wherefore no association between object classes is required across videos. Also the set of videos is assumed to be visually very similar. Furthermore, one presented evaluation [5] remains qualitative and the other one uses synthetically generated sequences [16] that paste a foreground video into different backgrounds. There is still a big disconnect between the idea of video co-segmentation to the challenges presented in video data from the web or personal video collections.

Contribution : This paper establishes multi-class video co-segmentation as a well defined challenge. We propose an approach that considers real video data, where neither the global number of appearance classes nor the number of instances in each images is known. Our method is based on the first application of distant-dependent Chinese Restaurant Processes for video data in order to formulate a video segmentation prior. Finally, we present the first quantitative evaluation of this new task which is performed on real video sequences.

2. Related Work

The idea of spatio–temporal analysis and segmentation of video data [6, 23, 22] has seen several refinements over the last years. More sophisticate probabilistic models [9, 13, 19] and the combination with tracking and segmentation algorithms [3, 14, 7] have greatly improved the applicability of such models. We organize and discuss related approaches as follows:

Video Segmentation In [14] long term point trajectories based on dense optical flow are used to cluster the feature points into temporally consistent segmentations of moving objects in the video. Similarly, in [7] with introduction of probabilistic region trajectories, they proposed to use spatial–temporal clustering on trajectories based on motion. Although their methods provide plausible solutions on video segmentation tasks, they lack a global appearance model that would relate segments across videos for the video co-segmentation task.

Image/Object Co-Segmenation Object co-segmentation [21] was first introduced to segment a prominent object based on an image pair in which it both appears. This idea has seen several refinements and today's state-of-the-art in co-segmentation can handle multiple objects [10]. Similar to [10], we assume the objects are shared between videos, therefore co-segmentation can be encouraged but not enforced. However, these approaches only look at single frames and do not consider spatio-temporal structure and motion, which we incorporate in our generative model. Also we overcome the issue of choosing a particular number of classes by employing a non-parametric prior.

Bayesian Non-parameterics for Image and Video Analysis In terms of learning appearance models, we relate to the application of topic models in the image domain [17]. This work has been extended to handle also spatial information [24] as well as part notions in infinite mixture models [18] and motion [12]. Non of these models have presented a video segmentation prior or described a generative model for appearance classes across multiple videos. From the modeling aspect, our work is inspired by the image segmentation method based on distant dependent Dirichlet Process (ddCRP) [8]. We present a model that employs ddCRP in order to formulate video segmentation prior as well as learning appearance models together with the segmentation across multiple videos.

Video Co-Segmentation Recently, two initial attempts [16, 5] have been made to approach video co-segmentation with a binary foreground/background segmentation task. But this setting makes quite strong assumptions and eliminates the problem of associating segments of multiple classes across frames and videos. In contrast, our method is the first to advance to less structured videos of multiple objects. We define and address a multi-class video co-segmentation task. In addition, we provide the first quantitative evaluation of this task on real videos.

3. Generative Multi-Video Model

The goal of this paper is to perform segmentation across multiple videos where the segments should correspond to the objects and segments of the same object class are linked together within and across videos. As motivated above, video segmentation on each video independently can lead to ambiguities that only can be resolved by reasoning across sequences. In order to deal with this problem we approach video cosegmentation by a generative model where videos are linked by a global appearance model. In order to be able to deal with an unknown number of object classes and object instances in each video, we make use of non-parametric bayesian modeling based on Dirichlet Processes. In particular, we define a video segmentation prior that proposes contiguous segments of coherent motion by a distance dependent Chinese Restaurant Process (ddCRP) as well as an infinite mixture model for the global appearance classes based on a Chinese Restaurant Process (CRP) [15].

After describing our video representation, we give an overview of Chinese Restaurant Processes (CRP) and extension to distant dependent Chinese Restaurant Processes (ddCRP) [2]. The ddCRP will then be used to define a video segmentation prior. In oder to define a generative model across video we add another layer on top that links the videos with a shared appearance model.

3.1. Video Representation

Given a set of videos $\mathcal{V}$, we start by a superpixel segmentation for each frame within the sequence and represent the video as a collection of superpixels. For every video $v \in \mathcal{V}$, we denote its total number of superpixels by N_v, and describe each superpixel i by its appearance feature x_i, spatio-temporal location s_i and motion vector m_i.

3.2. Distance Dependent Chinese Restaurant Processes (ddCRP)

We briefly introduce the basic idea of CRP and its extension to ddCRP. CRP is an alternative representation of Dirichlet process model and it defines the following procedure. Imagine a restaurant with an infinite number of tables. A sequence of customers come enter the restaurant and sit at randomly chosen tables. The i-th customer sits down at a table with a probability that is proportional to how many customers are already sitting at that table or opens up a new table with a probability proportional to a hyperparameter. Their seating configuration represents a random partition also called *table assignments*. Thus CRP provides a flexible prior distribution over table assignments where the number of tables is potentially infinite. Since the table assignment of each customer just depends on the number of people sitting at each table and is independent of the other ones, the ordering of customers does not affect the distribution over partitions and therefore exchangeability holds.

While in some cases there are spatial or temporal dependencies between customers, the exchangeability does not hold any more, the generalized process allowing non-exchangeable distribution over partitions is needed. The ddCRP was proposed to offer an intuitive way for modeling non-exchangeability and dependency. The main difference between the CRP and ddCRP is that rather than directly linking customers to tables with table assignments, in ddCRP customers sit down with other customers according to the dependencies between them, which leads to *customer assignments*. Groups of customers sit together at a table only implicitly if they can be connected by traversing the customer assignments. Therefore the i-th customer sits with customer j with a probability inversely proportional to the distance d_{ij} between them or sits alone with a probability proportional to the hyperparameter α:

$$p(c_i = j|D, f, \alpha) \propto \begin{cases} f(d_{ij}) & j \neq i \\ \alpha & j = i \end{cases} \tag{1}$$

where c_i is the customer assignment for customer i and $f(d)$ is the decay function and D denotes the set of all distances between customers. The decay function f should be non-increasing, takes non-negative finite values, and satisfies $f(\infty) = 0$. It describes how distances between customers affect the probability of linking them together.

3.3. ddCRP Video Segmentation Prior

We use the ddCRP in order to define a video segmentation prior. Customers correspond now to superpixels and tables correspond to object instances. The distance measure D and decay function f is now composed of two parts: $\{D^s, f^s\}$ and $\{D^m, f^m\}$ where the former one comes from the spatio-temporal distance and the latter one from motion similarities between superpixels.

$$p(c_i = j|D, f, \alpha) \propto \begin{cases} f^s(d^s_{ij})f^m(d^m_{ij}) & j \neq i \\ \alpha & j = i \end{cases} \tag{2}$$

Before measuring the spatio-temporal distance, we first use the optical flow vectors gained from TV-L1 model [4] in each pair of adjacent frames to find the neighbouring superpixels along temporal axis. Then the spatio-temporal distance D^s between superpixels is defined as the number of hops [8] required to travel from one superpixel to another. For the motion distance D^m between superpixels, we simply use the euclidean distances between mean motion vectors of superpixels for the motion similarities. For f^s, we use the *window decay* $f(d) = [d < A]$ which determines the probabilities to link only with customers that are at most distance A away. For f^m, we use the *exponential decay* $f(d) = e^{\frac{-d}{B}}$ which decays the probability of linking to customers exponentially with the distance to the current one, where B is the parameter of decay width. With the decay functions f^s and f^m for both spatio-temporal and motion domains, we have defined a distribution over customer (superpixel) assignments which encourages to cluster nearby superpixels with similar motions thus to have contiguous segments in spatio-temporal and motion domains. In Figure 2 we show samples from this ddCRP video segmentation prior for different hyperparameters. The prior proposes segments having contiguous superpixels with similar motion.

3.4. Generative Multi-Video Model

In this section we formulate a probabilistic, generative model that links the videos by a global appearance model that is also non-parametric. We consider the following hierarchical generative procedure of multiple video sequences:

Videos consist of multiple global object classes with different appearances, and for every video there are arbitrary number of instances which are located at different locations and possibly move over time. As our model has a hierarchical structure of layers for global classes and local instances which is very similar to the idea of Hierarchical Dirichlet Process [20], we use the same metaphor of its Chinese restaurant franchise representation in our case: There is a restaurant franchise (set of videos) with a shared menu of dishes (object classes) across all restaurants (videos). At

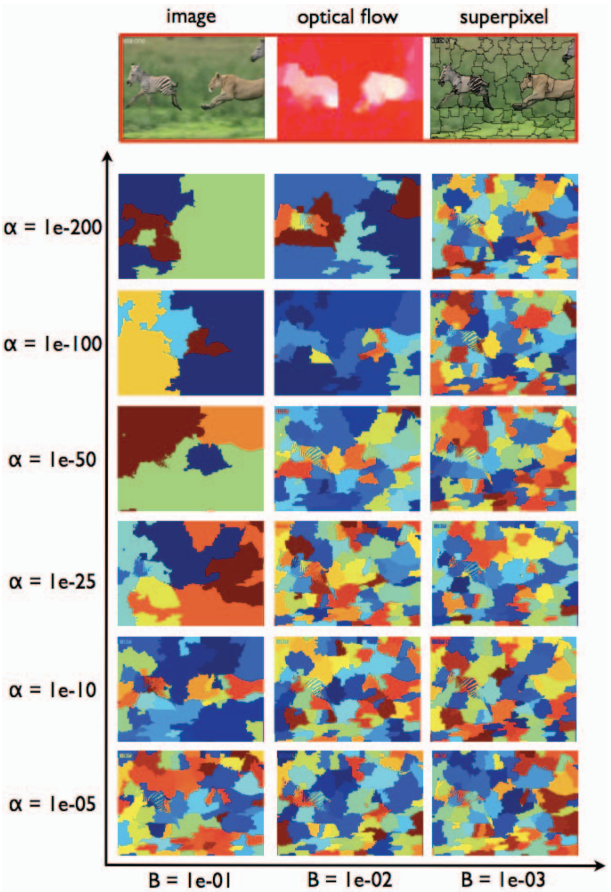

Figure 2. First Row (from left to right): Original image, motion map from optical flow, superpixel segmentation. Rest Rows: Samples from ddCRP video cosegmentation prior under different settings between concentration hyperparameter α and width parameter B for exponential decay function of motion f^m.

each table (object instance) of each restaurant one dish (object class) is ordered from the menu by the first customer (superpixel) who sits there, and it is shared among all customers (superpixels) who sit at that table (object instance). Multiple tables (object instances) in multiple restaurants (videos) can serve the same dish (object class). So the analogy is the following: restaurants correspond to videos, dishes correspond to object classes, tables correspond to instances, and customers correspond to superpixels. Here is a summary of the generative process:

1. For each superpixel i_v in video v, draw assignment $c_{i_v} \sim \text{ddCRP}(D, f, \alpha)$ to object instance
2. For each object instance t_v in video v, draw assignment $k_{t_v} \sim \text{CRP}(\gamma)$ to object class
3. For each object class k, draw parameters $\phi_k \sim G_0$
4. For each superpixel i_v in video v, draw observed feature $x_{i_v} \sim P(\cdot|\phi_{z_{i_v}})$, where $z_{i_v} = k_{t_{i_v}}$ the class assignment for i_v.

where G_0 is drawn from the DirichletProcess(γ, H_a) in order to define an infinite set of appearance models. H_a denote a Dirichlet prior on feature appearance distribution which is used as the base distribution for the process. γ is the concentration parameter for the Dirichlet process. For each global object class k discovered across video sequences, the parameter ϕ_k for its appearance model is sampled from G_0. We use a multinomial distribution η to describe the appearance model. Therefore given the observed appearance feature x_i for superpixel i, the likelihood of observed appearance feature for global object class k can be denoted as $p(x_i|\phi_k) = \eta_k(x_i)$.

Posterior Inference via Gibbs Sampling In order to incorporate the ddCRP video segmentation prior with the likelihood of superpixels to object instances whose appearance models are inherited from corresponding global object classes, we can now define a posterior distribution over customer assignments and use it to perform inference.

The goal of posterior inference is to compute posterior distribution for latent variables given observed data. The posterior for customer assignments is:

$$p(c_{1:N_v}|x_{1:N_v}, D, f, \alpha, \gamma) = \frac{\left(\prod_{i_v=1}^{N_v} p(c_{i_v}|D, f, \alpha)\right) p(x_{1:N_v}|z(c_{1:N_v}), \gamma)}{\sum_{c_{1:N_v}} \left(\prod_{i_v=1}^{N_v} p(c_{i_v}|D, f, \alpha)\right) p(x_{1:N_v}|z(c_{1:N_v}), \gamma)} \tag{3}$$

Here we use ddCRP $p(x_{1:N_v}|z(c_{1:N_v})$ as prior for all the possible customer configurations such that its combinatorial property makes the posterior to be intractable wherefore we use sampling techniques. As proposed in original ddCRP paper [2], Gibbs sampling is used where samples are iteratively drawn from the conditional distribution of each latent variable given the other latent variables and observations:

$$p(c_{i_v}|c_{-i_v}, x_{1:N_v}, D, f, \alpha, \gamma) \propto p(c_{i_v}|\alpha, D, f) \cdot p(x_{1:N_v}|z(c_{1:N_v}), \gamma) \tag{4}$$

The prior term is given in equation 2 and the likelihood term for multinomial appearance distribution is

$$\begin{aligned} p(x_{1:N_v}|z(c_{1:N_v}), \gamma) &= \prod_{l=1}^{|z(c_{1:N_v})|} p(x_{z(c_{1:N_v})=l}|z(c_{1:N_v}), \gamma) \\ &= \prod_{l=1}^{|z(c_{1:N_v})|} \eta_l(x_{z(c_{1:N_v})=l}) \end{aligned} \tag{5}$$

Resampling the global class (dish) assignment k follows typical Gibbs sampling method for Chinese Restaurant Process but consider all the features $x_{\mathcal{V}}$ and assignments $k^{\mathcal{V}}$

in the video set $\mathcal{V}$. The class assignment posterior of each table t_v in video v is:

$$p(k_{t_v} = l | k^{\mathcal{V}}_{-t_v}, x^{\mathcal{V}}, \gamma) \propto \begin{cases} m_l^{k^{\mathcal{V}}_{-t_v}} \eta_l^{k^{\mathcal{V}}_{-t_v}}(x_{t_v}) & \text{if } l \text{ is used} \\ \gamma \eta_l(x_{t_v}) & \text{if } l \text{ is new} \end{cases} \tag{6}$$

Here $k^{\mathcal{V}}_{-t_v}$ denotes the class assignments for all the tables in the video set $\mathcal{V}$ excluding table t_v, $x^{\mathcal{V}}$ is the appearance features of all superpixels within $\mathcal{V}$. Given the class assignment setting $k^{\mathcal{V}}_{-t_v}$, $m_l^{k^{\mathcal{V}}_{-t_v}}$ counts the number of tables linked to global class l whose appearance model is $\eta_l^{k^{\mathcal{V}}_{-t_v}}$. x_{t_v} stands for the appearance features of superpixels assigned to the table t_v.

3.5. Implementation Details

For computing the appearance feature representation for superpixels, we use the following pipeline: We use a similar procedure of dense patch extraction and patch description as in [10] in order to stay comparable to the image co-segmentation baseline which we will use in the experimental section. These patches are further quantized into a codebook of length 64 so that we can assign a color codeword to every image patch, which is based on a typical Bag-of-Words (BoW) image representation. Now we describe the appearance feature for each superpixel i by using the color codeword histogram x_i computed from the image patches whose center is located inside that superpixel.

For all our experiments we set the concentration parameter $\gamma = 1$ which is weakly informative. The hyperparameter on multinomial distribution for appearance information is assigned symmetric Dirichlet prior $H_a = \text{Dir}(2e + 2)$ which encourage to have bigger segments for global classes. The concentration parameter $\alpha = 1e - 100$ for the proposed video segmentation prior and the width parameter $B = 1e - 1$ for motion decay function f^m was determined by inspecting samples from the prior obtained from equation 2. We show examples in Figure 2 that displays the effect of the parameters. We set width parameter A for spatial decay function f^s to be 3 for all our experiments.

4. Experimental Results

In this section, we evaluate our generative video co-segmentation approach and compare it to baselines from image co-segmentation and video segmentation.

We first present our new dataset and the evaluation criterion that we propose. Then we present the results of our method and compare them to image co-segmentation and video segmentation baselines.

4.1. Dataset

We present a new Multi-Object Video Co-Segmentation (MOViCS) challenge, that is based on real videos and exposes several challenges encountered in online or consumer videos.

Up to now there is only a first attempt to propose a video co-segmentation benchmark [16]. The associated dataset is very limited as it only consists of one set of 4 videos that are synthetically generated. The same foreground video is pasted into 4 different backgrounds. Accordingly, their task is defined as binary foreground/background segmentation that does not address segmentation of multiple classes and how the segments are linked across videos by the classes.

In contrast to this early video co-segmentation approaches, we do not phrase the task as binary foreground/background segmentation problem but rather as a multi-class labeling problem. This change in task is crucial in order to make progress towards more unconstraint video settings as we encounter them on online resources and consumer media collections. Therefore, we propose a new video co-segmentation task of real videos with multiple objects in the scene. This makes a significantly more difficult problem, as not only object have to be correctly segmented but also assigned the same global class across video.

We propose the first benchmark for this task based on real video sequences download from youtube. The dataset has 4 different video sets including 11 videos with 514 frames in total, and we equidistantly sample 5 frames from each video that we provide ground truth for. Note that for each video set there are different numbers of common object classes appearing in each video sequence, and all the objects belonging to the same object class will be noted by the same label.

Unlike the popular image co-segmenation dataset *iCoseg* [1] which has similar lighting, image conditions and background or video segmentation dataset *moseg* [3] with significant motion patterns, our dataset exposes many of the difficulties encountered when processing less constraint sources. In Figure 3 we show examples of video frames for the four video sets together with the provided groundtruth annotations. Our sequences show different lighting conditions (e.g. tiger seq.), motion blur (e.g. chicken seq.), varying number of objects moving in and out (e.g. giraffe,elephant seq.), similar appearance between objects and background (e.g. tiger), etc. The MOViCS dataset and our code can be found at http://www.d2.mpi-inf.mpg.de/datasets.

4.2. Evaluation Metric

In order to quantify our results, we adopt the *intersection-over-union metric* that is also used in image co-segmentation tasks (e.g. [11]) as well as the PASCAL challenge.

$$M(S, G) = \frac{S \cap G}{S \cup G} \tag{7}$$

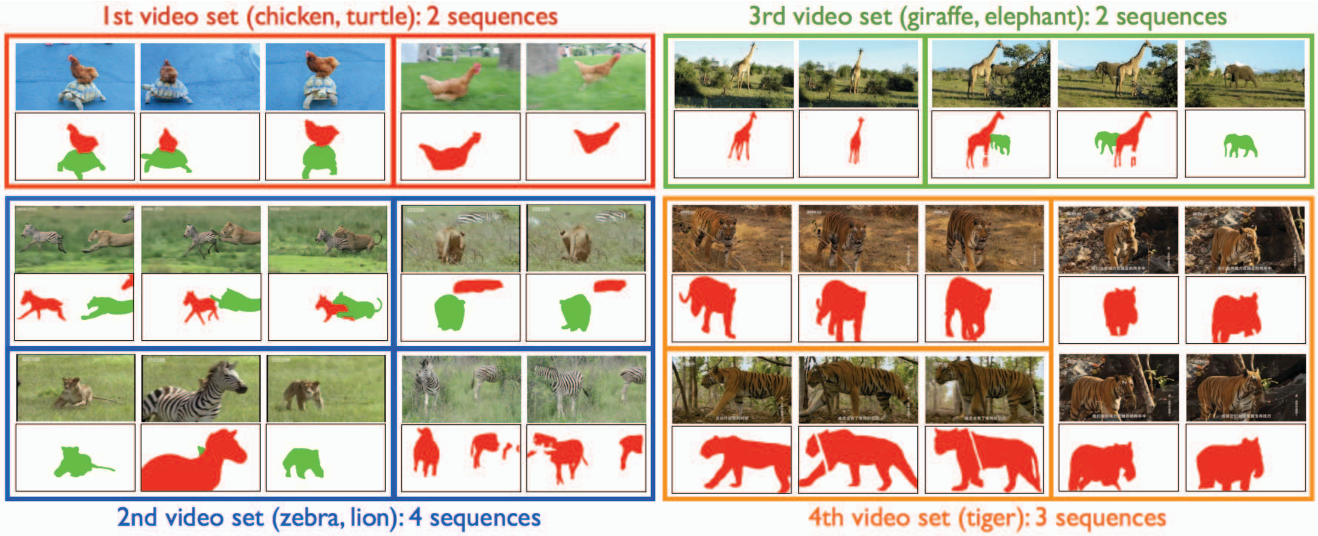
1st video set (chicken, turtle): 2 sequences
3rd video set (giraffe, elephant): 2 sequences
2nd video set (zebra, lion): 4 sequences
4th video set (tiger): 3 sequences

Figure 3. Summary of our proposed MOViCS dataset. Different color blocks stand for different video sets and the images within the same block come from the same video sequences.

where S is a set of segments and G are the groundtruth annotations.

We define our co-segmentation task as finding for each object class a set of segments that coincide with the object instances in the video frames. Therefore the algorithm has to group the segments by object class. We denote all segments grouped to an object class i by S_i. Therefore our evaluation assigns the object class to the best matching set of segments predicted by an algorithm:

$$\text{Score}_j = \max_i M(S_i, G_j) \tag{8}$$

Please note that this measure is not prone to over-segmentation, as only a single label is assigned per object class for the whole set of videos. We can further condense this performance measure into a single number by averaging over the classes.

$$\text{Score} = \frac{1}{C} \sum_j \text{Score}_j \tag{9}$$

where C is the number of object classes in the groundtruth.

Comparison to video segmentation A comparison to video segmentation methods is not straight forward. As each video is processed independently, there is no linking of segments across the videos. We therefore give the advantage to the video segmentation method that our evaluation links the segments across videos by the groundtruth.

4.3. Results

We evaluate our approach on the new MOViCS dataset and compare it to two state-of-the-art baselines from video segmentation and image co-segmentation. Our video segmentation baseline [14] is denoted by (VS) and the image co-segmentation baseline [10] is denoted by (ICS) whereas we use (VCS) for our video co-segmentation method. For both baselines we run the publicly available code of the authors on our data.

The performance numbers of the proposed method in comparison to the baselines are shown in Figure 4. With an overall performance of 48.75% of our method, we out perform VS by 22.08% and ICS by 31.5%.

Figure 7 shows a visualization of the results. First column is a frame of the video, second column shows the motion map, the third column shows the results of ICS, fourth column shows the result of VS and the last column shows the results of our VCS method.

Here the evaluation is performed per set of video sequences since the algorithm not only have to correctly segment the object instances but also link them to a consistent object class. As described in our evaluation metric, we don't allow for over-segmentation of the object classes in this experiment.

Also recall that VS doesn't have this property to link objects across videos. Therefore it has no notion of objects links across videos. As described above we give an advantage to the VS method by linking the segments across video via the groundtruth. Despite this advantage our method outperforms VS by a large margin for the first 3 video sets. Only on the tiger sequences VS performs better. It turns out that in this set the appearance is particularly hard to match across videos due to lighting and shadow effects, where the VS gets boosted by the additional information we had to provide for the comparison.

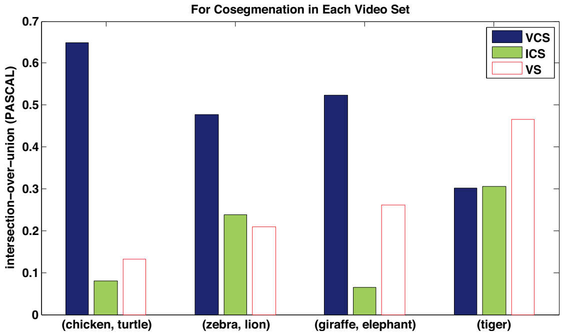

Figure 4. Comparison of co-segmentation accuracies between the our method (VCS), image co-segmentation (ICS) and video segmentation (VS) on the proposed MOViCS dataset. Only a single label is assigned per object class in the groundtruth for the whole set of videos.

Discussion The video segmentation baseline strongly depends on motion information in order to produce a good segmentation. When the motion map is noisy or there are objects moving together or with similar motion, segmentation errors occur. This issues are particular pronounced in the first video set where the chicken moves together with the turtle and the motion map is noisy due to fast motion in the second video. Our method handles such situations better and maintains a good segmentation despite the noisy motion information.

The image co-segmentation baseline has an assumption which expects a certain number of common object classes for all input images. This often cause problems for the less constraint settings that we are of interest in our study. For example in the second and third video sets in Figure 7, there are a varying number of objects moving in and out. The performance of image co-segmentation reduces in these settings. In addition, problems occur with wrongly merged object classes (lion with zebra, and giraffe with elephant). Our non-parametric approach seems to be better suited to deal with this variation on object instances and object classes and shows overall a more consistent segmentation.

Another interesting aspect of our model is how segmentation is supported by jointly considering all the videos of a set and learning a global object class model. Without this global appearance model, the performance decreases by 3.15% - still outperforming the baselines. We give an example in Figure 6 where the first row is the images from a single tiger video, the second row is the results by applying our proposed method only on this single sequence, and the last row is our VCS result while taking all videos in tiger set into account. We observe an improved segmentation that recovers parts of the tiger that were previously missing.

Analysis with over-segmentation In this analysis we relax the assumption that the sets of segments proposed by the method have to correspond to exactly one groundtruth object class each. Therefore, we now assign multiple set of segments to the same object class in the groundtruth. In Figure 5 we present the performance comparison under this relaxed setting. Please note that this relaxed measure doesn't penalize for non-existing links between the videos as well as over segmentation in the spatial domain. Overall, the performance improves, as over segmentation is not penalized. In average our method achieves a performance of 64.1% which still outperforms VS by 22.82% and ICS by 30.19%. The improvements under this measure are particular prominent on the video sets where appearance is hard to match across sequences. We take the fourth video set (tiger) as an example. In Figure 7 we observe that VS over-segments the tiger. This set of videos is challenging due to varying lighting conditions, shadows and appearance similarities with the background. Both ICS and VS do not match the object correctly across videos, as we can tell be the different coloring across videos. Our method does not show strong over-segmentation artifacts and also matches the object class across the first two videos.

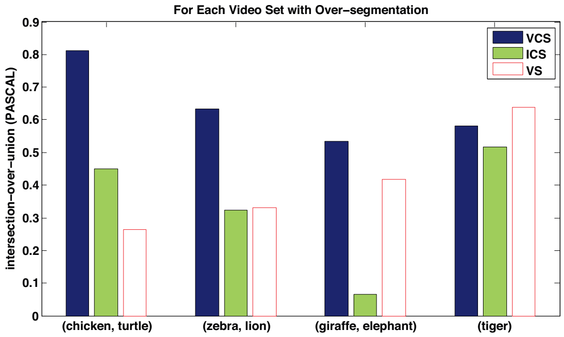

Figure 5. Comparison of co-segmentation accuracies between our approach (VCS) and baselines (ICS, VS) for MOViCS dataset. Allow over-segmentation which can assign multiple labels to the same object class in the groundtruth.

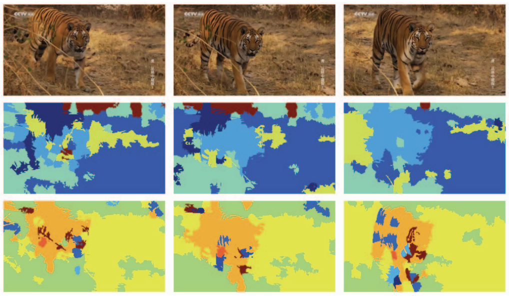

Figure 6. Example of improved results by segmenting across videos with a global object class model. First row: images from a single tiger video. Second row: results obtained from running our proposed method only on single tiger sequence. Third row: joint segmentation on all tiger videos.

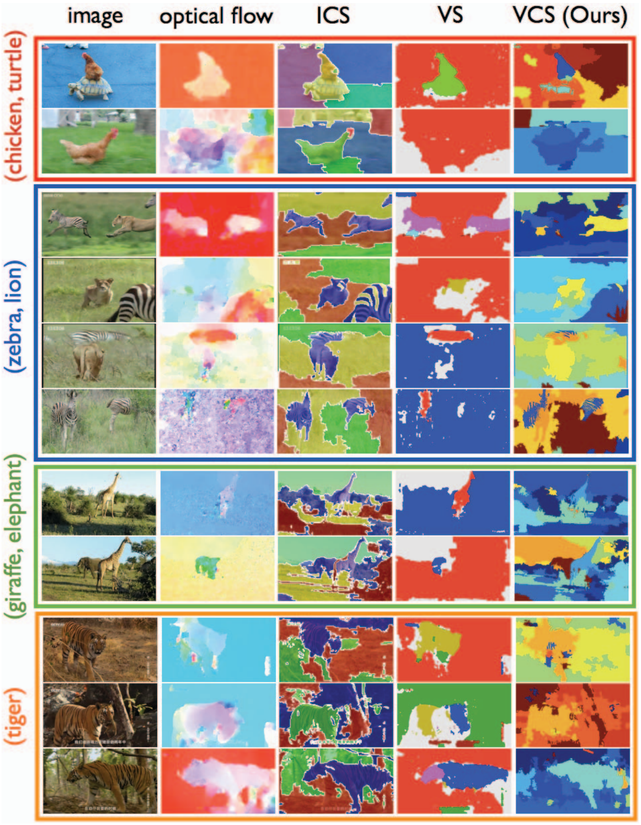

Figure 7. Examples of results from the proposed method (VCS) and baselines (ICS, VS) for all four video sets in MOViCS dataset.

5. Conclusion

We have proposed a non-parametric approach to the task of multi-class video co-segmentation. Our method incorporates a probabilistic video segmentation prior that proposes spatially contiguous segments of similar motion. We defined the first video co-segmentation challenge on multiple objects. The proposed Multi-Object Video Co-Segmentation (MOViCS) dataset is based on real videos and exposes challenges encountered in consumer or online video collections.

Our method outperforms state-of-the-art image co-segmentation and video segmentation baselines on this new task. We provide an analysis that give insights to the open challenges on this emerging task.

References

[1] D. Batra, A. Kowdle, D. Parikh, J. Luo, and T. Chen. icoseg: Interactive co-segmentation with intelligent scribble guidance. In *CVPR*, 2010. 5

[2] D. Blei and P. Frazier. Distance dependent chinese restaurant processes. In *ICML*, 2010. 2, 4

[3] T. Brox and J. Malik. Object segmentation by long term analysis of point trajectories. In *ECCV*, 2010. 2, 5

[4] A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145, 2011. 3

[5] D.-J. Chen, H.-T. Chen, and L.-W. Chang. Video object cosegmentation. In *ACM Multimedia*, 2012. 1, 2

[6] T. Darrell and A. Pentland. Robust estimation of a multilayered motion representation. In *IEEE Workshop on Visual Motion*, 1991. 2

[7] F. Galasso, M. Iwasaki, K. Nobori, and R. Cipolla. Spatio-temporal clustering of probabilistic region trajectories. In *ICCV*, 2011. 2

[8] S. Ghosh, A. Ungureanu, E. Sudderth, and D. Blei. Spatial distance dependent chinese restaurant processes for image segmentation. In *NIPS*, 2011. 2, 3

[9] H. Greenspan, J. Goldberger, and A. Mayer. A probabilistic framework for spatio-temporal video representation & indexing. In *ECCV*, 2002. 2

[10] A. Joulin, F. Bach, and J. Ponce. Multi-class cosegmentation. In *CVPR*, 2012. 2, 5, 6

[11] G. Kim and E. P. Xing. On multiple foreground cosegmentation. In *CVPR*, 2012. 5

[12] D. Kuettel, M. Breitenstein, L. Van Gool, and V. Ferrari. What's going on? discovering spatio-temporal dependencies in dynamic scenes. In *CVPR*, 2010. 2

[13] B. F. N. Jojic and A.Kannan. Learning appearance and transparency manifolds of occluded objects in layers. In *CVPR*, 2003. 2

[14] P. Ochs and T. Brox. Object segmentation in video: a hierarchical variational approach for turning point trajectories into dense regions. In *ICCV*, 2011. 2, 6

[15] J. Pitman. *Combinatorial stochastic processes*, volume 1875. Springer-Verlag, 2006. 2

[16] J. C. Rubio, J. Serrat, and A. M. López. Video cosegmentation. In *ACCV*, 2012. 1, 2, 5

[17] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman. Discovering objects and their location in images. In *ICCV*, 2005. 2

[18] E. Sudderth, A. Torralba, W. Freeman, and A. Willsky. Describing visual scenes using transformed objects and parts. *IJCV*, 2008. 2

[19] D. Sun, E. Sudderth, and M. Black. Layered image motion with explicit occlusions, temporal consistency, and depth ordering. In *NIPS*, 2010. 2

[20] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 2006. 3

[21] S. Vicente, C. Rother, and V. Kolmogorov. Object cosegmentation. In *CVPR*, 2011. 2

[22] J. Y. Wang and E. H. Adelson. Spatio-temporal segmentation of video data. In *SPIE*, 1994. 2

[23] J. Y. A. Wang and E. H. Adelson. Layered representation for motion analysis. In *CVPR*, 1993. 2

[24] X. Wang and E. Grimson. Spatial latent dirichlet allocation. *NIPS*, 2007. 2