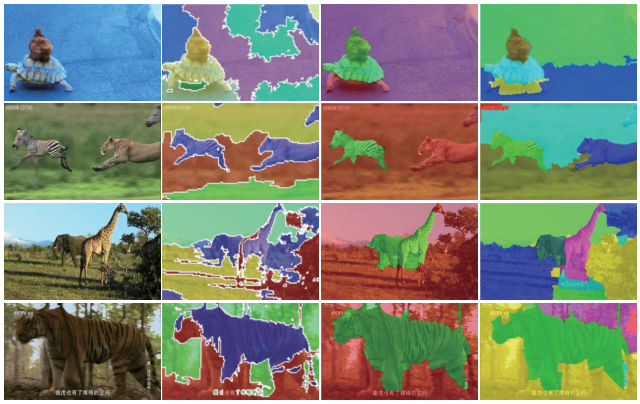
Semantic Single Video Segmentation with Robust Graph Representation*

Handong Zhao[1] **and Yun Fu**[1,2]
[1] Department of Electrical and Computer Engineering, Northeastern University, Boston, USA, 02115
[2] College of Computer and Information Science, Northeastern University, Boston, USA, 02115
{hdzhao,yunfu}@ece.neu.edu

Abstract

Graph-based video segmentation has demonstrated its influential impact from recent works. However, most of the existing approaches fail to make a semantic segmentation of the foreground objects, i.e. all the segmented objects are treated as one class. In this paper, we propose an approach to semantically segment the multi-class foreground objects from a single video sequence. To achieve this, we firstly generate a set of proposals for each frame and score them based on motion and appearance features. With these scores, the similarities between each proposal are measured. To tackle the vulnerability of the graph-based model, low-rank representation with $l_{2,1}$-norm regularizer outlier detection is proposed to discover the intrinsic structure among proposals. With the "clean" graph representation, objects of different classes are more likely to be grouped into separated clusters. Two open public datasets MOViCS and ObMiC are used for evaluation under both intersection-over-union and F-measure metrics. The superior results compared with the state-of-the-arts demonstrate the effectiveness of the proposed method.



(a) Input videos (b) ICS (c) VS (d) Ours

Figure 1: Semantic segmentation examples on dataset MOViCS. From top to bottom, the video samples are "chicken and turtle", "lion and zebra", "giraffe and elephant" and "tiger", respectively. From left to right, each column indicates (a) Input videos, (b) results from image cosegmentation method (ICS) [Joulin et al., 2012], (c) results from video segmentation method (VS) [Zhang et al., 2013] and (d) results from our proposed method. Different colors denote different labels.

1 Introduction

Video resource has grown tremendously with the development of digital technology. YouTube has over 100 hours of video uploaded per minute, so it is challenging to handle such big video data. Video object segmentation is one of the attracting applications for recent years, as it automatically generates a pixel-level boundary of foreground objects which benefits a lot for other higher-level applications, such as scene classification or understanding.

Generally the existing video segmentation methods can be categorized into the following groups: (1) graph-based model [Grundmann et al., 2010, Lee et al., 2011, Galasso et al., 2014,Zhang et al., 2013], (2) trajectory-based model [Li et al., 2013,Ochs and Brox, 2011,Palou and Salembier, 2013], (3) generative layered model [Galasso et al., 2012, Ma et al., 2013, Xu et al., 2013], and so on. It has been verified that the effectiveness of graph-based models enjoys the following advantages: (a) good generalizability to videos of arbitrary sizes, (b) rich mathematical supports for graph-model optimization, (c) encouraging segmentation result, and (d) less computational load [Galasso et al., 2012]. Thus in this work, we keep investigating the graph-based model.

Nevertheless, it is known that graph-based models are vulnerable to outliers. When the input data are heavily corrupted, graph-based models might fail [Candès et al., 2011, Liu et al., 2010]. For video segmentation, outliers and noises happen a lot, especially when the graph is built upon superpixels/supervoxels. It is because the current superpixel/supervoxel generation methods have difficulties in producing the exactly pixel-wise segmentation for objects. Af-

*This research is supported in part by the NSF CNS award 1314484, ONR award N00014-12-1-1028, ONR Young Investigator Award N00014-14-1-0484, and U.S. Army Research Office Young Investigator Award W911NF-14-1-0218.

ter over-segmenting video frames, those segments that cover both foreground object and background, but not stable across the frames, are considered as "outliers", e.g., the scroll caption in "tiger" case as shown in the bottom row of Figure 1. To alleviate the influence of outliers, several attempts have been made and encouraging results have been reported [Zhang et al., 2012, Zhang et al., 2013]. Zhang et al. [Zhang et al., 2012] represented each frame by a bunch of superpixels. Considering each superpixel as a vertex, the video sequence can be represented as a directed acyclic graph. The shortest path algorithm is used to select the graph vertices for all frames. However, this method cannot deal with the occlusion case. Most recently, Zhang et al. [Zhang et al., 2013] solved this problem by designing a dynamic programming algorithm which makes the graph model more robust. Although these attempts have demonstrated the effectiveness in handling outliers, they can only deal with one single foreground object case as shown in Figure 1(c).

To achieve the semantic segmentation, segment different objects based on their class information is challenging, especially for single video case. Grundmann et al. [Grundmann et al., 2010] proposed a hierarchical graph-based segmentation method, which over-segments a volumetric video graph into spatial-temporal regions (similar to superpixel/supervoxel) firstly. Then a region-graph is created. By finding the minimum spanning tree (MST) iteratively, the segmentation is achieved. It is deserving to notice that although different labels are assigned to different regions, the aim is to find regions with the same spatial-temporal information. While on opposite, we aim to produce the semantically meaningful segmentation for foreground objects. Similarly, another graph-based recent work [Galasso et al., 2014] by Galasso et al. focused on computation and memory costs but not semantic segmentation. For the majority of semantic segmentation works, multiple videos are needed. Chiu et al. [Chiu and Fritz, 2013] formulated this as a non-parametric bayesian model. Videos segmentation prior is introduced to leverage the segments of coherent motion and global appearance classes. Then objects with the same class property are linked across frames. Fu et al. [Fu et al., 2014] proposed a multi-stage selection graph method by leveraging the potential between two input videos and the potential within each video. Zhang et al. [Zhang et al., 2014] designed a regulated maximum weight clique extraction scheme. It balances well the spatially saliency and temporal consistency in object proposal (superpixels) tracklet selection step. The similar shape and appearance could be iteratively extracted by weighted groupings of objects. Unfortunately, these semantic segmentation methods require more than one video as input, which limits the scope of real-world applications.

In this work, to overcome the limitations of multiple videos as input and sensitivity to outliers, we fully exploit the spatial and temporal information in a single input video. As a trade-off, we assume that (1) the location of foreground objects in the successive frame has a relatively small shift and (2) multiple objects lie in separated latent subspaces. By considering the spatial consistency of foreground objects, object-like regions can be aligned across frames. With the usage of object appearance and motion, different objects can be clustered into different groups. In general, our method can be divided into four major steps as shown in Figure 2(b-e). First, a pool of scored "object-like regions" (proposals) for each frame are generated based on appearance and motion features. Second, to perform a robust graph representation, we conduct outlier detection with the help of low-rank representation. Third, spectral clustering is performed based on the clean representation. Finally, objects in the missing frame are recovered using spatial consistency. In general, we summarize our contributions as:

- To the best of our knowledge, this is the pioneer work to achieve semantic video segmentation using one single video, opposed to using multiple videos, required by other graph-based video segmentation methods.
- Low-rank representation based outlier analysis is conducted to uncover the intrinsic multiple structure of proposals. By characterizing the outliers via $l_{2,1}$-norm regularizer, this outlier detection technique can be applied in any graph-based segmentation methods.
- By fully exploiting the motion directions of foreground objects, we theoretically demonstrate the effectiveness of our proposed proposal objectness prediction method, which can well handle the inconsistent moving situation.

2 Related Work

In this section, we are talking about the related works of two major techniques used in the proposed method, i.e. graph-based video segmentation method and low-rank graph representation.

Graph-based object segmentation is a relatively new topic. Different from traditional video segmentation method [Shi and Malik, 1998], which works on individual pixel, the recent graph-based methods [Lee et al., 2011, Jain and Latecki, 2012, Zhang et al., 2013, Grundmann et al., 2010, Galasso et al., 2014] work on superpixel or supervoxel generated using object-like region extraction methods [Endres and Hoiem, 2010, Alexe et al., 2010]. This has an advantage in terms of processing speed. For instance, a 100×100 video clip has 100 superpixels per frame. The data to be processed with a pixel-based method are 1000 times more than with a superpixel-based method. Besides time saving, these object-like region extraction methods can provide a pool of semantic region candidates as a preprocessing step. Lee *et al.* [Lee et al., 2011] utilized a single cluster graph partitioning method [Olson et al., 2005] to discover the foreground object, then the authors refined the result by proposed pixel-level object labeling methods. Ma and Latecki [Jain and Latecki, 2012] treated the object region selection process as finding the maximum weight clique in a weight graph. Zhang *et al.* [Zhang et al., 2013] first constructed a directed acyclic graph based on object-like regions, then turned the region selection problem to find the maximal/minimal weight path problem. These methods that work well in segmenting one foreground object, however, will fail in multiple objects segmentation task.

Low-rank based graph representation attracts more and more attention because of its robustness to the data corruption [Candès et al., 2011, Liu et al., 2010]. Candès et al.

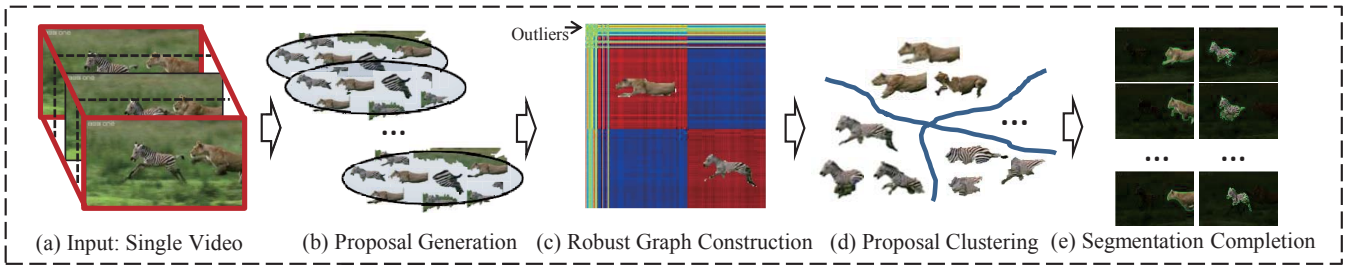


Figure 2: Framework of the proposed method is shown in the black dash box. To achieve the semantic video segmentation, only one single video is needed (a). Then we generate a pool of object-like regions ("Proposals") for each frame (b). To perform a robust graph representation, outlier detection with low-rank constraint is considered (c). With the new representation, spectral clustering is performed (d). To achieve a continuous object trajectory, we perform a post-processing procedure to align and recover the missing frames (e).

formulated the given observation matrix X by the sum of two terms, i.e. low-rank clean matrix X_0 and errors E, $X = X_0 + E$. This formulation implicitly assumes the underlying structure is single low-rank subspace. However, in real-world application, it is more likely that data are drawn from multiple subspaces. Thus, Liu et al. [Liu et al., 2010] proposed a low-rank representation (LRR) method to uncover the multiple subspace structure of data, which can represent the samples as linear combinations of the basis in a given dictionary, as $X = DZ+E$, where D is the "Dictionary", which can be set as data samples X itself in many applications [Liu et al., 2010]. Under the low-rank representation framework, $l_{2,1}$-norm regularization has demonstrated the effectiveness of handling outliers in many applications, for instance, face recognition [Shao et al., 2014], kinship verification [Xia et al., 2012], etc. This motivates us to characterize the video region outliers using $l_{2,1}$-norm regularizer.

3 Our Method

In this section, we illustrate our method in detail. Specifically, according to the order of video processing, object-like regions generation will be discussed firstly, then followed by the core part, proposal clustering with outlier detection. To complete the segmentation task, the post-processing step is followed at last.

3.1 Object-like Regions Generation

To achieve the multi-class foreground object segmentation task, the first thing is to segment all possible object-like regions (proposals). There exist several works focusing on finding proposals [Endres and Hoiem, 2010, Alexe et al., 2010]. However, these works are all single image based, which means the relationship between frames is not considered by simply applying these methods. We consider optical flow between successive frames to make a better prediction of objectness. Thus, the objectness of proposal can be calculated as:

$$S(r) = A(r) + M(r), \tag{1}$$

where r denotes a proposal, $A(\cdot)$ and $M(\cdot)$ stand for objectness scores measured by appearance feature and motion feature, respectively.

Specifically, we compute $A(r)$ using [Endres and Hoiem, 2010], which represents the objectness in the sense of appearance, such as color, texton, etc. Although it gives a reasonable likelihood measurement on proposal that is an object, it does not take motion into account since it is single-image based. With the help of consecutive frames at the input, objectness in the sense of motion $M(r)$ is measured using optical flow [Liu, 2009] as follows:

$$M(r) = 1 - \exp(-\xi(\mathbf{u}^r, \mathbf{v}^r)), \tag{2}$$

where $\mathbf{u}$ and $\mathbf{v}$ denote the optical flow calculated from current-to-previous (backward) direction, and current-to-next (forward) direction, respectively. $\mathbf{u}^r$ means the optical flow map in proposal r. Term $\mathbf{v}^r$ is defined in the same way. Since there are two directions x and y to quantify the optical flow, we define $\xi(\mathbf{u}, \mathbf{v})$ as:

$$\xi(\mathbf{u}, \mathbf{v}) = \frac{1}{\Omega}\|(\mathbf{u}_x \cdot \mathbf{v}_x) + (\mathbf{u}_y \cdot \mathbf{v}_y)\|_1, \tag{3}$$

where $\|\cdot\|_1$ is the sum of absolute value of all elements in the matrix [Lu et al., 2012]. It is worthy mentioning that the way we measure motion features within the proposal has a few benefits as follows: (1) It encourages the proposals with large motions by multiplication of forward and backward optical flows. Proposals with larger movements can be assigned to a high motion score; (2) Proposals with inconsistent moving direction will be depressed by this separate-direction motion scoring method. Without loss of generosity, let's assume (u_x, u_y) and (v_x, v_y) are the forward and backward optical flows on a single pixel. Only proposals with the similar direction (θ is small) are encouraged. For the extreme case, when $\theta = \pi/2$, and lengths (or norms) of two optical flows are the same, the motion score is zero. The effectiveness of this definition is based on the assumption that the foreground objects are moving consistently. Even object of interest moves erratically, the movement trajectory between frames are slight; (3) The noises induced by optical flow calculation or tiny background movement can be eliminated. Here an assumption that background objects' movements are trivial is made.

By normalizing two kinds of scores into the same scale, the quantitative measure on proposals is done by simply adding

two scores up. For the first and last frame, we only use appearance score due to the incapability of optical flow computation. To reduce the computational cost, 100 proposals with the highest scores in each frame are selected for the next step.

3.2 Proposal Clustering

Given the pool of scored object-like regions, our goal in this step is to group different regions into different categories based on class information. Dataset-oriented feature selection method can always achieve good results, however, it will narrow the application scope as well. To alleviate this dilemma, we only use two features in our method, color histogram for appearance and optical flow histogram for motion.

Firstly, we define color and optical flow similarity (distance) between two proposals r_m and r_n as follows:

$$\begin{aligned} X_c &= \exp(-\frac{1}{\mu_c}\chi^2_{\text{color}}(r_m, r_n)), \\ X_f &= \exp(-\frac{1}{\mu_f}\chi^2_{\text{flow}}(r_m, r_n)), \end{aligned} \tag{4}$$

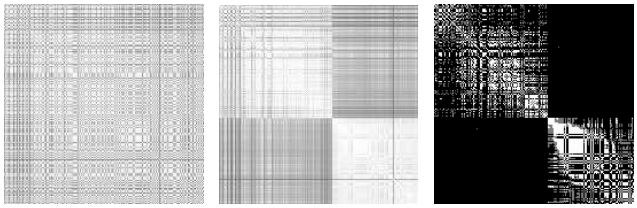
where $X_c \in \mathbb{R}^{d_c}$ and $X_f \in \mathbb{R}^{d_f}$ denote the χ^2-distance between color and optical flow histograms of r_m and r_n, respectively. μ_c and μ_f are the means of the all distances for color and optical flow histograms. Instead of combining them directly to form an affinity matrix, we make a simple but effective feature selection to determine whether or not to use the optical flow feature. The intuition is that the estimated optical flow map is not as robust as the color feature. Furthermore, the slight movement in the background could have a significant impact on finding foreground objects, especially when the foreground object has a relatively small movement or even no movement. Instead of giving the same weight on optical flow and color, we set a binary threshold λ to turn optical flow feature on/off.

Proposal Outlier Removal

The proposal generation method [Endres and Hoiem, 2010] used in the first step not only focuses on accurate object-like region prediction, but also focuses on region diversity, which means for each frame it gives a pool of object-like region candidates. So overlap between region candidates happens. In this case, the similarity between the overlapped proposals will have relatively high scores using our proposed measurement Eq. (4). It is the fact that the overlapped proposal covering both foreground object and background happens a lot, but not stable across the frames. We consider these proposals as "outliers". Obviously, removing these outliers will benefit the clustering result. Here, we concatenate two features X_c and X_f to get the $X = [X_c; X_f] \in \mathbb{R}^{d\times N}$, where $d = d_c + d_f$ denotes the new feature dimension for the outlier detection step, and N is the total number of proposals.

The intuition of low-rank representation is to find the new representation Z of original data X with intrinsic rank. Inspired by the fact that intrinsic rank of "clean" data is always small, we enforce the rank of Z in the objective following the previous works [Liu et al., 2010] as

$$\min_Z \text{rank}(Z), \;\; s.t. \;\; X = XZ. \tag{5}$$



(a) (b) (c)

Figure 3: Block property illustration with example case "chicken". (a) shows the similarity matrix before clustering (without grouping similar object together). Both (b) and (c) show the similarity matrix after clustering. (c) is the k-means result of (b), serving as the input for spectral clustering. Two blocks indicate cluster turtle and chicken respectively.

Recently, $l_{2,1}$-norm has been testified its efficiency on feature selection and outlier detection [Nie et al., 2010,Liu et al., 2010], with its definition as follows:

$$\|E\|_{2,1} = \sum_{j=1}^{n} \sqrt{\sum_{i=1}^{n} |E|_{ij}^2}. \tag{6}$$

Compared with l_p-norm ($p = 1, 2, \infty$), $l_{2,1}$-norm can well characterize the sample-specified outliers, since it is performed to select features across all points with joint sparsity. Then Eq. 5 can be rewritten as:

$$\min_{Z,E} \text{rank}(Z) + \beta\|E\|_{2,1}, \;\; s.t. \;\; X = XZ + E, \tag{7}$$

where β is the trade-off parameter balancing the two terms. Because of the discrete nature of the rank function, the Eq.(8) is difficult to solve. As a good surrogate of rank function, trace norm $\|\cdot\|_*$ is often used due to its convexity. Then the objective can be rewritten as:

$$\min_{Z,E} \|Z\|_* + \beta\|E\|_{2,1}, \;\; s.t. \;\; X = XZ + E. \tag{8}$$

The above convex optimization problem can be solved by several off-the-shelf methods. In this paper, exacted Augmented Lagrange Multiplier (ALM) method is applied [Bertsekas, 1982] to iteratively optimize Z and E.

After removing the outliers, we then define the similarity matrix W using the clean color feature $\tilde{X}_c$ and optical flow feature $\tilde{X}_f$ by

$$W = \tilde{X}_c^{\mathrm{T}}\tilde{X}_c + \lambda\tilde{X}_f^{\mathrm{T}}\tilde{X}_f, \tag{9}$$

where $\tilde{X}_c = X_c - E_c$ and $\tilde{X}_f = X_f - E_f$. λ is the binary threshold defined above. The similarity matrix W has a property of block-diagonal by putting the proposals within the same cluster together, as shown in Figure 3. Proposals within the cluster have higher similarity (white); the correlation between two clusters has a lower similarity (black). The sizes of two blocks are not the same. In fact, the perfect clustering result should have the same size in this case, as the video sequence shown in Figure 3 is "chicken" from dataset MOViCS, and both chicken and turtle exist in all frames. Thus, by selecting one best proposal for each frame, both of the block sizes should be equal to the frame number. However, this

ideal case rarely happens due to the following reasons: (1) there always exist some similar proposals (for most cases, they are parts of the foreground objects) considered as inliers. In this "chicken" case, the lower right block represents chicken, the upper left block represents turtle. As we can see clearly from Figure 3(c), the two block sizes are not the same, representing the different number of proposal "chicken" and "turtle".

As we obtain the similarity matrix W, to speed up the clustering step, k-nearest neighbour algorithm is used before performing spectral clustering. By applying the basic graph knowledge, the normalized Laplacian matrix L is defined as $L = I_N - D^{-1/2}WD^{-1/2}$, where I_N is the identity matrix with N-dimension and degree matrix D is the diagonal degree matrix with its elements determined by the corresponding row sums of W.

With L, we can achieve the semantic segmentation with multiple classes. For simplicity, we illustrate our idea on the case class number k equals 2. For a given proposal y, we consider it as foreground when y is set to 1 and as background when set to -1. Thus, our problem can be formulated as,

$$\min \mathbf{y}^{\mathrm{T}} L \mathbf{y}, \quad \text{s.t. } \mathbf{y}^{\mathrm{T}} D^{1/2} \mathbf{1}_N = 0,\ \mathbf{y} = \{-1, 1\}^N, \quad (10)$$

where $\mathbf{1}_N$ denotes the N-dimensional vector of all ones. Following the classical spectral clustering methods [Ng et al., 2001]. This objective function can be solved by using singular value decomposition (SVD).

4 Experimental Result

In this section, we make a quantitative evaluation on the proposed method and other baselines. Since we focus on graph-based model, the state-of-the-arts of graph-based image cosegmentation (ICS) [Joulin et al., 2012] and graph-based video segmentation (VS) [Zhang et al., 2013] are selected as baselines. The evaluation is performed on two public video databases containing multiple foreground objects. To demonstrate the effectiveness of the outlier detection part, we consider two versions of our method, i.e. **Ours-1** denotes our model without outlier detection, and **Ours-2** denotes our model with outlier detection.

4.1 Dataset

In order to evaluate the performance of multi-class foreground object segmentation, many traditional video segmentation databases are not suitable, because they mainly focus on one moving object segmentation task, such as database SegTrack [Tsai et al., 2010], GaTech [Grundmann et al., 2010] or Persons and Cars Segmentation Database [Zhang et al., 2013]. However recently, video cosegmentation works provide the source databases for evaluating the semantic segmentation, since these works focus on foreground segmentation with different object labels.

In this work, we select two open public datasets, MOViCS [Chiu and Fritz, 2013] and ObMiC [Fu et al., 2014], where MOViCS is the first benchmark for multi-class video cosegmentation task. It has 4 different video cases with a total of 11 video clips, including "chicken", "lion", "giraffe", and "tiger". All the foreground objects with the same class are assigned by the same label.

Table 1: Results of intersection-over-union metric on datasets MOViCS and ObMiC.

Datasets		ICS	VS	Ours-1	Ours-2
MOViCS	chicken	0.467	0.518	0.605	**0.654**
	lion	0.596	0.405	0.613	**0.646**
	giraffe	0.419	0.139	0.422	**0.441**
	tiger	0.424	0.669	0.688	**0.698**
ObMiC	dog	0.362	0.294	0.418	**0.454**
	person	0.459	0.168	0.477	**0.516**
	monster	0.509	0.139	0.546	**0.570**
	skating	0.138	0.123	0.144	**0.211**
Average		0.422	0.307	0.489	**0.524**

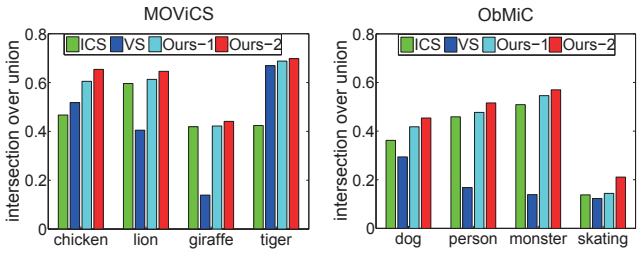


Figure 4: Experimental results of intersection-over-union metric on dataset MOViCS (left) and ObMiC (right).

Dataset ObMiC also contains 4 sets of videos. Each has two video clips. Different from MOViCS only containing animals, ObMiC extends the object to people and cartoon figures. Compared with MOViCS, ObMiC is more challenging due to the following reasons: (1) scale, most foreground objects in MOViCS are relatively large (except "chicken" case), which is less sensitive to the bad proposal generation; (2) Slight motion, objects in ObMiC have less movements, which might result in the failure of optical flow feature.

4.2 Comparison

As the graph-based video segmentation method, one of our baselines is the state-of-the-art video segmentation from Zhang et al. [Zhang et al., 2013]. In addition, ICS can be tailored to the video segmentation work without taking the spacial-temporal information into account. Thus, we also make a comparison with the state-of-the-art ICS [Joulin et al., 2012]. Specifically, an important parameter of ICS which needs tuning is the class number. In the experiments, this number is set in the range of [4, 8]. In VS, two key parameters of GMM model and MRF model are tuned in the range of [0.1, 2]. For all methods, we run the publicly available code and report the best performance.

To make a quantitative evaluation, we use "intersection-over-union" metric used in PASCAL challenge and previous works [Chiu and Fritz, 2013]. It is well defined to measure the segment region on both precision and recall rate as,

$$M(S, G) = \frac{S \cap G}{S \cup G}, \quad (11)$$

where S stands for a set of segments and G is the groundtruth annotation. Following the common clustering performance

Table 2: Results of F-measure on datasets MOViCS and ObMiC.

Datasets		ICS	VS	Ours-1	Ours-2
MOViCS	chicken	0.625	0.649	0.703	**0.751**
	lion	0.727	0.567	0.740	**0.765**
	giraffe	0.582	0.415	0.621	**0.635**
	tiger	0.608	0.731	0.807	**0.845**
ObMiC	dog	0.571	0.505	0.609	**0.647**
	person	0.663	0.395	0.671	**0.709**
	monster	0.710	0.363	0.750	**0.773**
	skating	0.393	0.349	0.362	**0.434**
Average		0.610	0.497	0.658	**0.695**

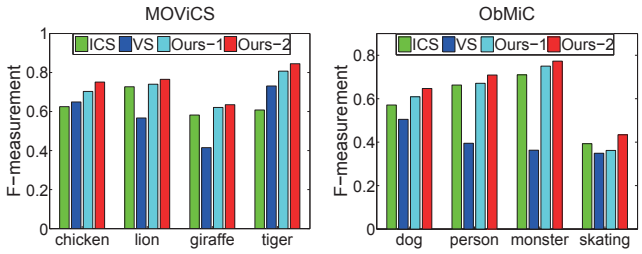


Figure 5: Experimental results of F-measure on dataset MOViCS (left) and ObMiC (right).

measurement criteria, we assign the class label to the segments with the best matching to groundtruth.

We present the comparison result of intersection-over-union metric on both datasets in Figure 4. The corresponding visualizations of all methods are shown in Figure 1 and Figure 6, respectively. By observation, several conclusions can be drawn: (1) Our-2 performs the best. (2) Compared with Our-1, Our-2 with outlier detection has a modest improvement, indicating the effectiveness of the low-rank based robust graph construction part. (3) For most multi-class objects cases, i.e. "lion and zebra" and "giraffe and elephant" in MOViCS and all the cases in ObMiC, ICS performs better than VS. This is because VS fails to consider the multi-class objects scenario, which is a common scene in real-world. Take "giraffe and elephant" as example in Figure 1, VS cuts both the giraffe and elephant out with same label. Moreover, similar to many other previous video segmentation methods which only segment one foreground object, VS only gives the results of "zebra" in "zebra and lion" case. (4) All the methods perform bad for the case "skating" in ObMiC. Two possible reasons may explain this: (a) the foreground objects (skaters) are relatively small, which easily results in a large false positive; (b) The female skater has different appearances for her upper body and lower body, which is difficult to segment. The statistics of performance measured by intersection-over-union metric are tabulated in Table 1. We highlight the highest score using bold font. It is deserved to note that for MOViCS, our method improves the performance of each case by 26.25%, 8.39%, 5.25% and 4.33%, respectively. For ObMiC, we boost the performance by 25.41%, 12.42%, 11.98% and 52.9%, respectively. On average, the performance bar is raised by 24.17%.

Note that although intersection-over-union metric is a good measurement on segmentation results, it is weighted more by recall than precision. To make a wholesome evaluation, F-measure is used as a complementary measurement. Compared with the groundtruth, the F-measure with different parameters γ is calculated as:

$$F_\beta = \frac{\left(1+\gamma^2\right) \text{Precision} \times \text{Recall}}{\gamma^2 \times \text{Precision} + \text{Recall}}. \tag{12}$$

We set $\gamma^2 = 0.3$ following Achanta et al. [Achanta et al., 2009] to weigh precision more than recall. Figure 5 shows the comparison results between our method and other methods. We can observe the same trend showing that our method performs the best. The exact F-measure result numbers are shown in Table 2. On average, our method outperforms ICS by 13.93% and VS by 39.84%, respectively.

5 Conclusions

We have proposed a semantic graph-based video segmentation method by clustering the objects of different classes into separate groups. To make a robust graph representation, we have analyzed the cause of outliers and modeled them with $l_{2,1}$-norm regularizer under low-rank constraint. With the "clean" representation, spectral clustering is used to get the different foreground object groups. The experimental results demonstrate that our method outperforms state-of-the-art image cosegmentation and graph-based video segmentation baselines on the existing datasets MOViCS and ObMiC.

References

[Achanta et al., 2009] Achanta, R., Hemami, S. S., Estrada, F. J., and Süsstrunk, S. (2009). Frequency-tuned salient region detection. In *CVPR*, pages 1597–1604.

[Alexe et al., 2010] Alexe, B., Deselaers, T., and Ferrari, V. (2010). What is an object? In *CVPR*, pages 73–80.

[Bertsekas, 1982] Bertsekas, D. (1982). *Constrained Optimization and Lagrange Multiplier Methods*. Athena Scientific.

[Candès et al., 2011] Candès, E. J., Li, X., Ma, Y., and Wright, J. (2011). Robust principal component analysis? *J. ACM*, 58(3):11.

[Chiu and Fritz, 2013] Chiu, W.-C. and Fritz, M. (2013). Multi-class video co-segmentation with a generative multi-video model. In *CVPR*.

[Endres and Hoiem, 2010] Endres, I. and Hoiem, D. (2010). Category independent object proposals. In *ECCV*, pages 575–588.

[Fu et al., 2014] Fu, H., Xu, D., Zhang, B., and Lin, S. (2014). Object-based multiple foreground video co-segmentation. In *CVPR*.

[Galasso et al., 2012] Galasso, F., Cipolla, R., and Schiele, B. (2012). Video segmentation with superpixels. In *ACCV*, pages 760–774.

[Galasso et al., 2014] Galasso, F., Keuper, M., Brox, T., and Schiele, B. (2014). Spectral graph reduction for efficient image and streaming video segmentation. In *CVPR*, pages 49–56.

[Grundmann et al., 2010] Grundmann, M., Kwatra, V., Han, M., and Essa, I. A. (2010). Efficient hierarchical graph-based video segmentation. In *CVPR*, pages 2141–2148.

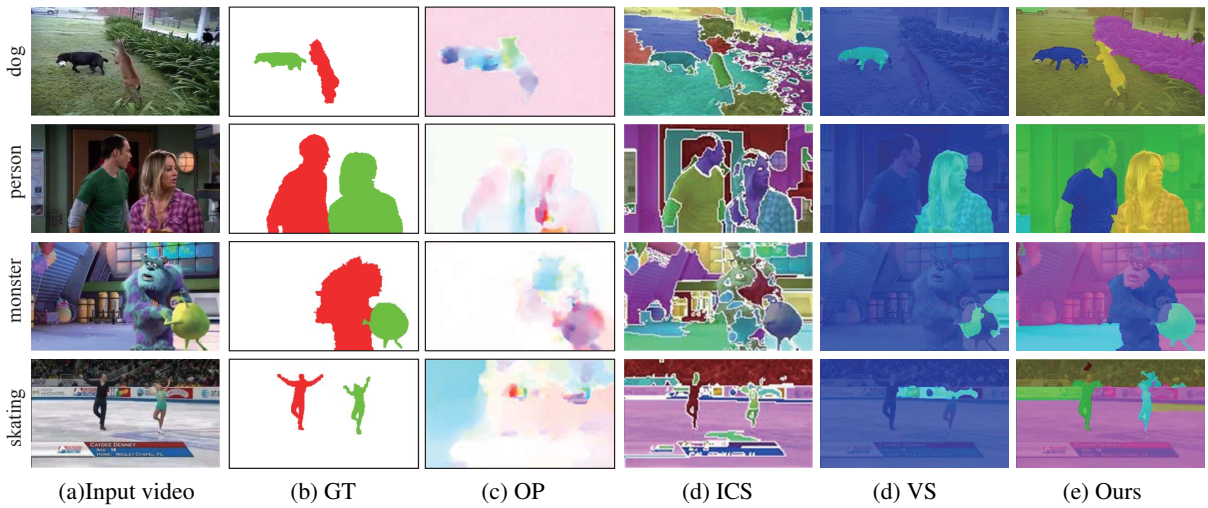


Figure 6: Comparison results on ObMiC. From top to bottom, each row shows a video cases, i.e. dog, person, monster, skating. The first three columns represent input video, ground-truth (GT) and optical flow (OP). Other columns from left to right show the comparison results by ICS [Joulin et al., 2012], VS [Zhang et al., 2013] and the proposed method with outlier detection. Different colors denote different labels.

[Jain and Latecki, 2012] Jain, A. and Latecki, L. J. (2012). Maximum weight cliques with mutex constraints for video object segmentation. In *CVPR*, pages 670–677.

[Joulin et al., 2012] Joulin, A., Bach, F., and Ponce, J. (2012). Multi-class cosegmentation. In *CVPR*, pages 542–549.

[Lee et al., 2011] Lee, Y. J., Kim, J., and Grauman, K. (2011). Key-segments for video object segmentation. In *ICCV*, pages 1995–2002.

[Li et al., 2013] Li, F., Kim, T., Humayun, A., Tsai, D., and Rehg, J. M. (2013). Video segmentation by tracking many figure-ground segments. In *ICCV*.

[Liu, 2009] Liu, C. (2009). *Beyond Pixels: Exploring New Representations and Applications for Motion Analysis*. PhD thesis, Massachusetts Institute of Technology.

[Liu et al., 2010] Liu, G., Lin, Z., and Yu, Y. (2010). Robust subspace segmentation by low-rank representation. In *ICML*, pages 663–670.

[Lu et al., 2012] Lu, C.-Y., Min, H., Zhao, Z.-Q., Zhu, L., Huang, D.-S., and Yan, S. (2012). Robust and efficient subspace segmentation via least squares regression. In *ECCV*, pages 347–360.

[Ma et al., 2013] Ma, T., Chatterjee, S., and Vidal, R. (2013). Coarse-to-fine semantic video segmentation using supervoxel trees. In *ICCV*.

[Ng et al., 2001] Ng, A. Y., Jordan, M. I., and Weiss, Y. (2001). On spectral clustering: Analysis and an algorithm. In *NIPS*, pages 849–856.

[Nie et al., 2010] Nie, F., Huang, H., Cai, X., and Ding, C. H. Q. (2010). Efficient and robust feature selection via joint $l2,1$-norms minimization. In *NIPS*, pages 1813–1821.

[Ochs and Brox, 2011] Ochs, P. and Brox, T. (2011). Object segmentation in video: A hierarchical variational approach for turning point trajectories into dense regions. In *ICCV*, pages 1583–1590.

[Olson et al., 2005] Olson, E., Walter, M., Teller, S. J., and Leonard, J. J. (2005). Single-cluster spectral graph partitioning for robotics applications. In *Robotics: Science and Systems*, pages 265–272.

[Palou and Salembier, 2013] Palou, G. and Salembier, P. (2013). Hierarchical video representation with trajectory binary partition tree. In *CVPR*, pages 2099–2106.

[Shao et al., 2014] Shao, M., Kit, D., and Fu, Y. (2014). Generalized transfer subspace learning through low-rank constraint. *IJCV*, 109(1-2):74–93.

[Shi and Malik, 1998] Shi, J. and Malik, J. (1998). Motion segmentation and tracking using normalized cuts. In *ICCV*, pages 1154–1160.

[Tsai et al., 2010] Tsai, D., Flagg, M., and M.Rehg, J. (2010). Motion coherent tracking with multi-label mrf optimization. *BMVC*.

[Xia et al., 2012] Xia, S., Shao, M., Luo, J., and Fu, Y. (2012). Understanding kin relationships in a photo. *TMM*, 14(4):1046–1056.

[Xu et al., 2013] Xu, C., Whitt, S., and Corso, J. J. (2013). Flattening supervoxel hierarchies by the uniform entropy slice. In *ICCV*.

[Zhang et al., 2012] Zhang, B., Zhao, H., and Cao, X. (2012). Video object segmentation with shortest path. In *ACM MM*, pages 801–804.

[Zhang et al., 2013] Zhang, D., Javed, O., and Shah, M. (2013). Video object segmentation through spatially accurate and temporally dense extraction of primary object regions. In *CVPR*, pages 628–635.

[Zhang et al., 2014] Zhang, D., Javed, O., and Shah, M. (2014). Video object co-segmentation by regulated maximum weight cliques. In *ECCV*, pages 551–566.