Video Segmentation via Multiple Granularity Analysis

Rui Yang[†], Bingbing Ni[†], Chao Ma[‡], Yi Xu[†], Xiaokang Yang[†]
[†]Shanghai Jiao Tong University [‡]The University of Adelaide
[†]{yangrui,nibingbing,xuyi,xkyang}@sjtu.edu.cn,[‡]c.ma@adelaide.edu.au

Abstract

We introduce a Multiple Granularity Analysis framework for video segmentation in a coarse-to-fine manner. We cast video segmentation as a spatio-temporal superpixel labeling problem. Benefited from the bounding volume provided by off-the-shelf object trackers, we estimate the foreground/background super-pixel labeling using the spatio-temporal multiple instance learning algorithm to obtain coarse foreground/background separation within the volume. We further refine the segmentation mask in the pixel level using the graph-cut model. Extensive experiments on benchmark video datasets demonstrate the superior performance of the proposed video segmentation algorithm.

1. Introduction

Video segmentation aims at separating target objects of interest from noisy background, and has received considerable attention with a wide range of computer vision applications, such as 3D reconstruction [34], video summarization [11], etc. Numerous algorithms have been proposed during the past decade with focus on developing graphical models, e.g., Markov Random Field (MRF), and Conditional Random Field (CRF), to estimate target motions for each pixel (optical flow) [6, 30, 5] or superpixel [22, 41]. Despite their favorable performance in several datasets, video segmentation still faces two main challenges. First, when graphical models are leveraged to compute temporal consistency in the pixel or superpixel level, there often exist mismatching pairs between consecutive frames. For example, the supervoxel algorithm [15, 44, 38] models the temporal consistency using superpixels for each frame. The inaccuracy caused by the mismatching of superpixels is inevitably aggregated frame by frame, and finally leads video segmentation algorithms to fail. We also note that developing a superpixel model across several frames is computationally inefficient. Second, object level motions estimated by visual tracking algorithms often contain noisy background as tracking results in the form of bounding boxes are not tightly around target objects. Video segmentation benefits little from the recent progress of visual tracking algorithms [24, 28].

To address these challenges, we present a novel framework of applying the multiple instance learning (MIL) algorithm [8] to both spatial and temporal domains for video segmentation. In contrast to most machine learning algorithms that assign every training instance with a label, MIL assigns bags of instances with labels. In the binary case, a bag is labeled positive if at least one instance in that bag is positive, and the bag is labeled negative if all the instances in it are negative. MIL is able to classify instances with missing or noisy labels based on the labeled bags as training data. This motivates us to apply the MIL algorithm to compute the temporal consistency in the temporal domain. For example, temporal adjacent and similar superpixels always belongs to the same label (i.e, foreground or background), since motion between consecutive frames can not be too significant. On the other hand, object level motions estimated by visual tracking algorithms in the form of bounding boxes provide rich information for the video segmentation task despite partial noisy background inside bounding boxes. Built on state-of-the-art tracking algorithms, we properly enlarge the tracked bounding boxes to meet the requirement of applying MIL as in [42]. We find that MIL deals with the noisy background well and provides an accurate envelop of the true foreground object masks. This significantly facilitates video segmentation.

Similar to [42], we use superpixels [1] as instances for learning the spatio-temporal MIL algorithm. Since MIL often benefits from more discriminative features, we propose a multi-scale CNN feature based descriptor to strengthen the discriminative power of each superpixel. To obtain better segmentation in the temporal domain, we use the tracking results from the state-of-the-art tracker [24, 28] to construct positive and negative bags of superpixels. To make full use of temporal consistency, we take the spatio-temporal consistency into consideration and use the superpixels over a short temporal span to construct bags. To initialize the labels of bags, we use a superpixel clustering paradigm by grouping superpixels with similar features. In addition, once we have the spatio-temporal segmentation

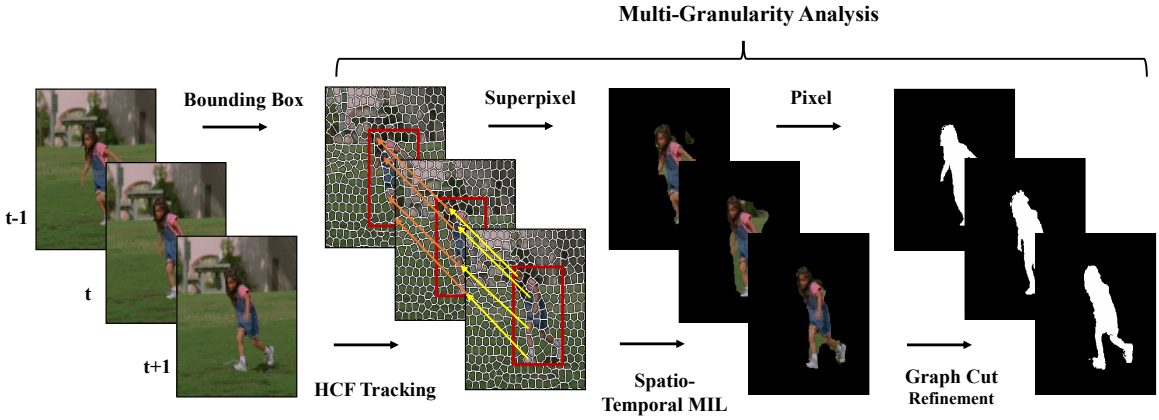

Figure 1. Overview of the proposed model. For segmentation, we first adopt state-of-art tracking method to generate self-adapting size bounding boxes to transfer segmentation task into weakly-supervised problem. We them build a spatio-temporal MIL model by extending structural information per image to several frames. Upon obtaining this coarse superpixel-level segmentation results, we apply graph-cut method to refine boundaries on a pixel level.

masks from MIL, we apply the pixel-level graph cut algorithm [20, 31] to refine the segmentation results utilizing spatio-temporal consistency cue. Therefore we can regard the proposed method as a **multi-granularity framework** for video segmentation problem which can effectively segment target objects from the background in a coarse to fine fashion. In the coarsest level (object), off-the-shelf object tracker is applied to the whole video sequence, yielding a candidate volume of object bounding boxes. In the middle level (superpixel), we perform multiple instance learning within the candidate volume to obtain a coarse segmentation result. In the finest level (pixel), segmentation mask is further refined via graph cut like algorithm. We comprehensively evaluate our algorithm on two popular video segmentation datasets, the Segtrack 2.0 [22] and Davis Dataset [32] released in CVPR 2016. The results demonstrate the superiority of our video segmentation method over the state-of-the-art algorithms.

Our contributions are three-fold:

1. We propose a novel video segmentation framework by applying the multiple instance learning in both the temporal and spatial domains to deal with the issues of temporal superpixel mismatching and noisy background within tracked bounding boxes, respectively. To the best of our knowledge, this is the first attempt to use MIL for video segmentation.

2. We explore multiple levels of information to segment target objects from the background in a coarse-to-fine manner. Tracking results provide object-level candidate bounding boxes. Superpixel is the middle level abstraction of target appearance. We also apply the graph cut algorithm to refine the segmentation mask in the pixel level.

3. The proposed algorithm significantly advances state-of-the-art video segmentation algorithm on public benchmark datasets with large-scale videos.

2. Related Works

Video Object Segmentation. A large number of methods have been proposed for resolving the video segmentation problem. Several works [12, 14] aim at annotating each pixel in every frame. Others focus on separating one or several objects from the background [37, 29, 29, 20, 22, 25, 36].

Existing algorithms [4, 16, 39, 31] widely use graphical models to merge similar or adjacent superpixels or pixels. Work of Galasso *et al*. [13] developed a graphic model based on spectral clustering. Grundmann *et al*. [15] proposed a greedy agglomerative clustering method. In [46], Yi *et al*. used Markov Random Field for unconstrained video segmentation that depends on tight integration of multiple cues. Work of Khoreva *et al*. [19] treated the video segmentation problem by highlighting the importance of classifiers and features. While Jang *et al*. [17] used MRF as an optimization method. These works utilize graphical models to maintain temporal consistency between a few frames and refine the outline of each segment. Tsai *et al*. [38] added a pairwise potential term for applying the graph cut method. This object-flow paradigm considers two consecutive frames at one time, and cannot handle the noisy mismatching issues over a long temporal span. It is noticeable that existing algorithms benefit little from the progress of visual tracking, that adapts to target appearance changes and provides tracked bounding boxes with the object informa-

tion.

Multiple Instance Learning. Multiple Instance Learning has been proposed with several variances, including D-D [27], EM-DD [49], since Dietterich *et al*. [8] primitively introduced this method in 1997. Since then, MIL has attracted considerable attention in computer vision. Voila *et al*. [47] introduced Mil-boost with purpose of combining the Adaboost algorithm [35] with the MIL paradigm. MIL shows favorable performance on object detection tasks. Recently, MIL successfully pushed forward online visual tracking [48, 3], saliency detection [47] and image retrieval [23]. Most of these works focus on establishing a robust way of updating an appearance model by training MIL in one or several frames. In these tasks, MIL demonstrates the robustness to input instances with ambiguous labels and achieves satisfying performance. Wu *et al*. [42] proposed a MILcut method for the task of interactive image segmentation, where users input a bounding box to initialize the segmentation task. None of these works attempted to apply MIL to the video segmentation tasks.

In our work, we utilize the MIL method on packs of consecutive frames, which obtains better temporal consistency over time as well as strengthens the classification ability from instances (superpixels) with noisy labels. Different from previous works [48, 3], we assign bags of superpixels with positive or negative labels using the sweeping line paradigm and consider the spatio-temporal relationship.

3. Methodology

3.1. Motivation and Overview

To address the challenging video segmentation problem, we propose a multiple granularity analysis framework to tackle this problem in a coarse-to-fine manner, as illustrated in Figure 1. In the coarsest level, we apply off-the-shelf visual trackers to obtain temporally continuous foreground object bounding boxes throughout frames, resulting in candidate foreground volumes (bounding volume). In the middle level, a multiple instance learning algorithm is applied onto the superpixels inside the bounding volume to identify the set of foreground superpixels (coarse segmentation), by exploring spatio-temporal consistency. In the finest level, a graph-cut based algorithm is applied on the pixels of this coarse segmentation, which yields the final segmentation result. We will introduce details about these three processing components in the following.

3.2. Coarse Granularity Analysis: Bounding Volume Generation

Recently, deep learning based techniques have significantly improved the performance of visual tracking on benchmark dataset [43]. Tracking results in the form of bounding boxes provide rich object information. To sequentially obtain an outline of target objects inside bounding boxes, we utilize the recently proposed HCFT tracker [24] that learns adaptive correlation filters over deep CNN features to handle significant appearance changes. We additionally enhance the HCFT tracker with a scale estimation module to get tighter bounding boxes. To ensure that tracked bounding boxes provide sufficient and correct information for following process, we set up two scaling factor c_p and c_n during implementation. Inside each bounding box in each frame, we can always find target objects regardless partial noisy background. While outside the bounding boxes are entirely the background.

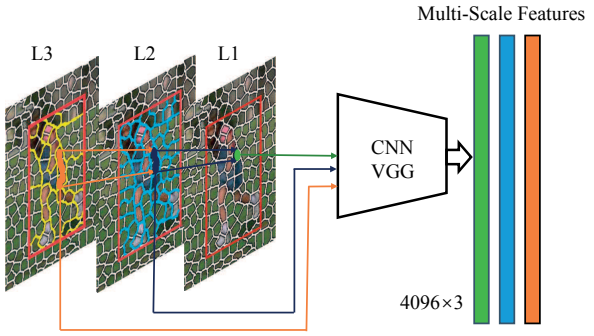

Figure 2. Multi-scale CNN superpixel feature. Superpixels in L1, generated by SLIC, are merged into larger superpixels by clustering, which contains richer topological information. Superpixels of different sizes are then fed into VGG network to predict multiple-layered CNN features.

3.3. Middle Granularity Analysis: Superpixel Parsing

Once the sequence of object bounding boxes (i.e., bounding volume) is obtained, we can first decompose the foreground regions into superpixels and cast the video segmentation problem as identifying foreground superpixel across all frames. Superpixel labeling based approach has achieved great success recently in image segmentation, however, directly extending the image based algorithm towards video domain is not feasible. Three issues need to be addressed. First, as video contains larger variations than image, it is more challenging to separate foreground superpixels from background ones in videos. To this end, we propose a multi-scale CNN descriptor (feature representation) for encoding each superpixel for further processing, which inherits the good property of robustness and discriminating power from multi-scale analysis and CNN. Second, the appearance of an object/superpixel is usually consistent across nearby temporal frames and this temporal consistency property should be fully explored to enhance segmentation. We therefore propose video based MIL algorithm to effectively recognize foreground superpixels. The two innovations are explained in detail as follows.

Multi-scale CNN Superpixel Feature Sizes and numbers of grains define the clustering ability of superpixels. While dense and small superpixels show more sensitivity to detailed color and texture changes, they may not contain as much structural information as bigger superpixels do. Therefore based on this consideration, we build a coarse-to-fine hieratical model for superpixel feature representation.

More specifically, we first generate small-sized dense superpixels for each frame using SLIC [1] method, and each superpixel containing 100-150 pixels. Then we extract RGB as well as DCNN features for each superpixel. We feed each frame into VGG-19 net, and upsample the output to the frame size, then averagely pool pixels to get deep features for each superpixel. To build multi-scale superpixel representation, we start from the finest over-segmentation. For each superpixel, and we gather its locally connected and similarly looked neighbors to merge into large superpixels. We define the similarity between superpixels by Euclidean distance of feature vectors. Here, the larger superpixel could be considered as its irregular context window. This process is iterated for several times Q. We then concatenate the DCNN features extracted from the superpixels of each level into a feature representation vector by dimension of $4099 \times Q$ with 3 dimensions RGB and 4096 dimensions CNN features. The multi-scale superpixel CNN feature extraction process is illustrated in Figure 2.

Video-based Multiple Instance Learning Now each bounding volume could be regarded as a bag of superpixels, and the segmentation task can be defined as inferring the foreground/background label for each superpixel contained in the bounding volume. According to the previous works in image segmentation, we can cast this label inferring problem as a multiple instance learning problem.

Multiple Instance Learning requires at least one positive instance in the positive bag and no positive instance in the negative bag. These requirements can be naturally met by exploring the tightness of bounding boxes [21]. Suppose we have a bounding box B for object P in image G, we define the top, bottom, left and right side of the bounding box as B_a,B_b,B_c,B_d. If,

$$(G \backslash B) \cap P = \emptyset,$$

$$P \cap B_a \neq \emptyset, P \cap B_b \neq \emptyset, P \cap B_c \neq \emptyset, P \cap B_d \neq \emptyset$$

we can describe this bounding box as effective and tight: the bounding box covers the object completely and every side of the box intersects the outline of the object. By virtue of such properties, we sample the bounding-box area with horizontal or vertical rectangular slices with their lengths or heights equal to that of the bounding boxes, and every sample contains at least one positive instance while the superpixels outside the bounding box can all be considered as negative instances. As proved by Wu *et al*. [42], when the objects inside are continuous, MIL constraints can be met on a single image. For video segmentation task, we extend this method on a pack of several consecutive frames, *i.e*. training positive bags and negative bags on this pack of frames together. Efficient tracking algorithms could export accurate bounding box with self-adapting scale. These algorithms employ region-proposal or edge-detection method to approximate bounding boxes to the edge of foreground object, enforcing bounding boxes to tightly and efficiently contain complete foreground object and a little background noises. By this means, we build a spatio-temporal MIL model, which is robust to errors due to the weak-supervision nature of MIL. Also, it favors temporal consistency since it's applicable on a larger pack of frames than supervoxel. Topological information within each frame is considered in our method as well.

The detailed mathematic formulation of the multiple instance learning based video segmentation algorithm (superpixel labeling) is presented as follows. For training a MIL model, suppose we have N bags on a pack of images including K consecutive frames, where $N = \sum_{k=1}^{K} N_k$. In ith bag, feature vector $X_i = \{x_{i1}, ..., x_{ij}, ..., x_{im}\}$ represents the features of M instances within, $y_i \subset \{0, 1\}$ for bag label and y_{ij} for the unknown instance labels. Therefore training data possess a form of $\{(X_1, y_1), ..., (X_n, y_n)\}$ which indicates feature vector and unknown label of instances. We define p_i as the possibility that ith bag is positive and p_{ij} for jth instance inside. For conserving the topological information in each frame as well as temporal consistency between frames, we construct our loss function as follows:

$$L(\phi) = L_t(\phi) + \lambda L_s(\phi) \tag{1}$$

where $L_t(\phi)$ indicates temporal term and it stands for the negative log-likelihood of bags in one pack of frames. $L_s(\phi)$ is spatial term which enforces connectivity constrained within single frames. ϕ indicates weak classifiers, which will be explained in below.

Definitions for these two terms are as follows:

$$L_t(\phi) = -log \prod_i p_i^{y_i} (1 - p_i)^{(1-y_i)} \tag{2}$$

$$L_s(\phi) = \sum_{k=1}^{K} \sum_{i=1}^{N_k} \sum_{f(\alpha,\beta)} \rho_{\alpha\beta} \|p_{\alpha i} - p_{\beta i}\|^2 \tag{3}$$

In spatial term, $f(\alpha, \beta)$ refers to the pairs of adjacent instances contained in one bag, and $\rho_{\alpha\beta}$ represents the length of their common boundary. Under this condition, the neighbouring superpixels tend to share similar labels.

For temporal model $L_t(\phi)$, defining $\phi(x_{ij})$ as an instance-level weak classifier, we employ Adaboost framework to combine weak classifiers into a strong one Φx_{ij}. θ as the weight of each weak-classifier and R as number of

weak classifiers selected, strong classifier can be expressed as:

$$\Phi(x_{ij}) = \sum_{r=1}^{R} \theta_r \phi_r(ij) \tag{4}$$

After computing the response of each weak classifier, strong classifier is formed by selecting the classifiers with greatest discriminative capacity. The possibility of each instance is given by softmax function:

$$p_{ij} = \frac{1}{(1 + \exp(-y_{ij}))} \tag{5}$$

Using ISR-Boost [18] likelihood ratio, the bag possibility can be deducted by:

$$S_i = \sum_{j \subset i} \exp(y_{ij}) \tag{6}$$

$$p_i = \frac{S_i}{(1 + S_i)} \tag{7}$$

where S is an internal quantity defined in ISR-Boost method.

Following MIL-boosting method [45], the weight on instance for Eqn. (2) equals the derivative of loss function with respects to change in y_{ij}, and we apply chain rule on it:

$$\omega_{ij}^t = \frac{\partial log L_t(\phi)}{\partial y_{ij}} = \frac{\partial log L_t(\phi)}{\partial p_i} \frac{\partial p_i}{\partial p_{ij}} \frac{\partial p_{ij}}{\partial y_{ij}} \tag{8}$$

In our case,

$$\begin{aligned} \omega_{ij} = \omega_{ij}^t + \omega_{ij}^s &= \frac{\partial log L(\phi)}{\partial y_{ij}} \\ &= \frac{\partial log L_t(\phi)}{\partial y_{ij}} + \lambda \frac{\partial log L_s(\phi)}{\partial y_{ij}} \\ &= \frac{\partial log L_t(\phi)}{\partial p_i} \frac{\partial p_i}{\partial p_{ij}} \frac{\partial p_{ij}}{\partial y_{ij}} + \lambda \frac{\partial log L_s(\phi)}{\partial p_{ij}} \frac{\partial p_{ij}}{\partial y_{ij}} \end{aligned} \tag{9}$$

according to ISR-boost,

$$\frac{\partial p_i}{\partial p_{ij}} = (\frac{1 - p_i}{1 - p_{ij}})^2 \tag{10}$$

and

$$\frac{\partial log L_s(\phi)}{\partial y_{ij}} = \sum_{f(\alpha,\beta)} 2\rho_{\alpha\beta}(p_{\alpha i} - p_{\beta i}) \tag{11}$$

The remaining parts in Eqn. (9) can be deducted using simple derivation of Eqn. (5) and (2).

The goal of optimization process is to find the best strong classifiers that separate foreground superpixels and background. Each iteration for this optimization contains four steps: calculating the weight, training weak classifiers with this weight, minimizing lost function and updating strong classifiers. The number of total weak classifiers is R', which equals to 200 in our case.

3.4. Fine Granularity Analysis: Pixel-wise Refinement

In the previous section, a spatio-temporal MIL method is applied on superpixels to address video segmentation problem. However, as the MIL algorithm is usually based on the bag defined by the tracked bounding box which is often inaccurate on object boundaries, the output of the MIL tends to bias some background superpixels to the label of "foreground". We therefore require a post-processing step to refine and smooth the outline of the foreground object created by the MIL segmentation algorithm operated on superpixel level. To this end, we propose to apply a graph cut style algorithm for this refinement purpose, which combines the information provided by pixels and superpixels of the estimated region in order to build a multi-level, coarse-to-fine refinement model. Note that segmentation results from the previous step are used to initialize this refinement process. Following Tsai *et al*. [38], the mathematic form of the cost function for the refinement process is defined as:

$$E_{total} = E_{pixel}(\mu) + E_{sp}(\nu) + E_{pairwise}(\mu, \nu) \tag{12}$$

where E denotes energy function for each frame in the video, which could be further expanded as energy function for pixel, superpixel and pairwise term. μ and ν represent pixels and superpixels respectively. In pixel term, we use GMM model on RGB and SVM on CNN features, while in superpixel term, RGB feature as well as superpixel clustering CNN feature are fed in energy function. $E_{pairwise}$ is the term that considers the compatibility between pixel and superpixel, which can be expressed as:

$$E_{pairwise}(\mu, \nu) = \begin{cases} 1 - |p(\mu_i) - p(\nu_j)| & \text{if different labels} \\ 0 & \text{else} \end{cases} \tag{13}$$

where we enforce the pixels and superpixels with similar features to have same labels. For each pixel/superpixel, we construct its neighbor across three frames, i.e., consider both within frame spatial consistency and between-frame temporal consistency.

4. Experiments

4.1. Implementation Details

We first use tracking method based on HCF-tracking [24] and KCFDP-tracking [7] to generate self-adapting bounding boxes for all frames, and decompose every frame with SLIC-superpixel algorithm. For an image of 854 $\times$ 480 image, around 3200 superpixels are produced. To generate bags for MIL method on superpixel-level, we shrink the box by 5% ($c_p = 0.05$) to sample positive bags and expand the box by 15% to sample negative ones ($c_n = 0.15$) , assuring the tightness and effectiveness of the box. To be

Table 1. Segmentation results (%) on Segtrack v2 dataset. Accuracy are presented by overlap ratio. Mean per sequence results are calculated by average performance of objects appeared in the same sequence. Mean per objects accuracy is the mean value for all objects results listed on the table.

Sequence	[38]	[22]	[41]	[20]	[40]	Ours
Online ?	✓		✓		✓	✓
Girl	87.9	**89.2**	83.7	87.7	52.4	87.2
Birdfall	57.4	62.5	**77.5**	49	32.5	55.2
Parachute	94.5	93.4	94.4	**96.3**	69.9	94.6
Cheetah-Deer	33.8	37.3	**63.1**	44.5	33.1	45.2
Cheetah-Cheetah	**70.4**	40.9	35.3	11.7	14	68.9
Monkeydog-M	54.4	71.3	**82.2**	74.3	22.1	59.7
Monkeydog-D	53.3	18.9	21.1	4.9	10.2	**59.2**
Penguin-1	**93.9**	51.5	92.7	12.6	20.8	92.3
Penguin-2	87.1	76.5	**91.8**	11.3	20.8	88.2
Penguin-3	89.3	75.2	**91.9**	11.3	10.3	91
Penguin-4	88.6	57.8	**90.3**	7.7	13	85.2
Penguin-5	80.9	66.7	76.3	4.2	18.9	**82.6**
Penguin-6	85.6	50.2	**88.7**	8.5	32.3	87.8
Drifting-1	84.3	74.8	67.3	63.7	43.5	**85.6**
Drifting-2	39	60.6	**63.7**	30.1	11.6	38.5
Hummingbird-1	69	54.4	58.3	46.3	28.8	**70.3**
Hummingbird-2	72.9	72.3	50.7	**74**	45.9	73.3
BMX-Person	88	85.4	88.9	87.4	27.9	**93.7**
BMX-Bike	7	24.9	5.7	**38.6**	6	9.5
Frog	81.4	72.3	61.9	0	45.2	**83.8**
Worm	**89.6**	82.8	76.5	84.4	27.4	87.6
Soldier	**86.4**	83.8	81.1	66.6	43	85.5
Monkey	88.6	84.8	86	79	61.7	**90.2**
Bird of Paradise	95.2	94	93	92.2	44.3	**96.3**
Mean /Object	74.1	65.9	71.8	45.3	30.7	**75.5**
Mean /Sequence	76.4	71.4	72.7	58.1	37.7	**77.6**

noted, negative bags are sampled near the expanded bounding box so as to represent well near-object background. We set number of weak classifiers $r = 15$ out of $R' = 200$ classifiers, and $\lambda = 0.05$, and generalized mean is used as the softmax model with exponent set to 1.5. When bounding boxes take the size larger than 2/3 of whole image, we directly cast whole image as positive bags.

For multi-scale CNN feature extractor, we extract the 3th, 6th, 10th, 14th and 18th layers of VGG network and do two iterations of superpixels-clustering. On pixel level, we feed each frame into VGG-19 and then upsample the output to frame size to obtain CNN features at each pixel position. Moreover, the weight for CNN and color feature for pixels and superpixels are respectively 3, 1, 5, 1, while general weight for pixels and superpixels are 1 and 15 to leverage their difference in numbers. All these parameters are fixed during the experiments for all datasets.

4.2. Results and Discussion

We conduct experiments on two popular datasets.

Table 2. Segmentation results (%) on Davis dataset. Accuracy presented by overlap ratio. * stands for incomplete video name.

Sequence	[2]	[10]	[26]	[33]	[31]	[9]	Ours
bear	93.7	92.9	**95.5**	90.6	89.8	90.7	92.9
blackswan	87.1	93	94.3	90.8	73.2	87.5	**94.7**
bmx-bumps	49	33.6	43.4	30	24.1	**63.5**	52.8
bmx-trees	47.3	22.9	38.2	24.8	18	21.2	**64.2**
boat	61.9	**70.5**	64.4	61.3	36.1	0.7	63.1
breakdance	**71.3**	47.8	50	56.7	46.7	67.3	59.4
breakdance-fla*	73.3	43	72.7	72.3	61.6	**80.4**	73.1
bus	74.9	66.8	86.3	83.2	82.5	62.9	**87.5**
camel	**79.5**	64	66.9	73.4	56.2	76.8	69.6
car-roundabout	78.6	72.6	85.1	71.7	80.8	50.9	**88.8**
car-shadow	70.1	64.5	57.8	72.3	69.8	64.5	**82.4**
car-turn	86.5	83.4	84.4	72.4	85.1	83.3	**86.6**
cows	81.1	75.6	**89.5**	81.2	79.1	88.3	82.2
dance-jump	47	49	74.5	52.2	59.8	71.8	**75.2**
dance-twirl	**64.4**	44.4	49.2	47.1	45.3	34.7	54.3
dog	62.1	67.3	72.3	77.4	70.8	**80.9**	64.9
dog-agility	66.3	69.9	34.5	45.3	28	65.2	**70.3**
drift-chicane	**80.6**	24.3	3.3	45.7	66.7	32.4	62.3
drift-straight	**75.3**	61.8	40.2	66.8	68.3	47.3	55.8
drift-turn	**85.6**	71.7	29.9	60.6	53.3	15.4	68.3
elephant	68.6	75	85	65.5	82.4	51.8	**87.8**
flamingo	85	53	**88.1**	71.7	81.7	53.9	83.6
goat	64.1	**73.1**	66.1	67.7	55.4	1	58.4
hike	90	66.4	75.5	87.4	88.9	91.8	**93.2**
hockey	**77.5**	67.7	82.9	64.7	46.7	81	62.2
horsejump-hi*	64.9	58.6	80.1	67.6	57.8	83.4	**86.5**
horsejump-lo*	54.5	66.3	60.1	60.7	52.6	65.1	**79.8**
kite-surf	**65.4**	50	42.5	57.7	27.2	45.3	59.4
kite-walk	73.6	50.9	**87**	68.2	64.9	81.3	86.4
libby	65.5	29.5	77.6	31.6	50.7	63.5	**85.2**
lucia	82	83.6	**90.1**	80.1	64.4	87.6	82.1
mallard-fly	**79.9**	53.6	60.6	54.1	60.1	61.7	65.2
mallard-water	75.5	75.1	90.7	68.7	8.7	76.1	**91.6**
motocross-bu*	**82.7**	76.1	40.1	30.6	61.7	61.4	69.9
motocross-jum*	**76**	58.3	34.1	51.1	60.2	25.1	64.4
motorbike	68.8	50.6	56.3	71.3	55.9	71.4	**82.8**
paragliding	87.7	95.1	87.5	86.6	72.5	88	**95.4**
paragliding-lau*	59.9	58.9	64	57.1	50.6	**62.8**	62.6
parkour	81.5	34.2	75.6	32.2	45.8	**90.1**	77.1
rhino	86.4	71.6	78.2	79.4	77.6	68.2	**89.4**
rollerblade	55.4	72.6	58.8	45	31.8	81.4	**89.7**
scooter-black	70.4	62.6	33.7	50.4	52.2	16.2	**72.5**
scooter-gray	65.3	12.3	50.8	48.3	32.5	58.7	**71.3**
soapbox	68	**75.8**	78.9	44.9	41	63.4	65.3
soccerball	85.6	9.7	84.4	82	84.3	82.9	**92.3**
stroller	60	65.6	76.7	59.7	58	**84.9**	30
surf	**94.4**	94.1	49.2	84.3	47.5	77.5	92.6
swing	70.9	11.5	78.4	64.8	43.1	85.1	**82.1**
tennis	71.4	76.5	73.7	62.3	38.8	**87.1**	58.9
train	53.5	87.3	87.2	84.1	83.1	72.9	**91.6**
Mean	72.4	60.7	66.5	63.1	57.5	64.1	**75.2**

Segtrack v2 Dataset. We evaluate our algorithm on the Segtrack v2 Dataset [22]. This dataset contains 14 videos

Figure 3. Demonstration results for our segmentation methods on Davis Dataset. The output mask contour is labeled in red color. Our method is capable of segmenting foreground objects under difficult situation such as deformation, motion blur, appearance change, and occlusions. Detailed information is also well preserved such as girl's hair in the third row and swan's tail. Better viewed in color.

Table 3. Detailed analysis for example sequences from Davis and Segtrack v2 dataset. Tracking precision is calculated by bounding-box overlap. MIL refers to segmentation masks computed after superpixel-MIL process without refinement. Pixel-level precision refers to final segmentation results.

Methods	Tracking (%)	MIL	Pixel-level
Girl	92.5%	70.1%	87.2%
Parachute	77.5%	74.2%	94.6%
Monkeydog(M)	69.0%	30%	59.7%
Blackswan	95.2%	73.2%	94.7%
Hike	96.3%	77.9%	93.2%
Stroller	27.6%	17.2%	30.0%

with 24 objects and 947 annotated frames. It includes various challenging videos with occlusion, motion blur, appearance change and deformation. Some of the videos contains multiple objects with interactions, which can be segmented in turn. Here, we present our results in Table 1.

Table 1 illustrates the mean accuracy and accuracy per object/sequence for the proposed algorithm as well as other state-of-art methods [38, 41, 20, 15, 40, 22]. The top preforming methods are shown in bold letter. The accuracy is represented by the overlap (IoU) of the predicted model and ground truth mask.

As shown in Table 1, the proposed algorithm outperforms existing methods on this dataset, especially for fast-moving, non-rigid objects with complex deformation, such as *BMX-Person, Monkey, Frog, Hummingbird*. Notwithstanding that these videos contain large deformations or clustered background changes, the proposed method achieves excellent performances. The online superpixel-tracking methods [38, 41, 15] do not perform well on these videos since even one misclassification will be propagated throughout the entire video to lower accuracy. However, in the proposed algorithm, we use MIL method to enhance robustness, which has been proven effective. Meanwhile, some methods [40, 15] does not consider pixel-level information. Their results are, thus, inaccurate on object boundaries. In the proposed algorithm, boundaries are refined by our multi-granularity system.

For videos with large appearance change, such as *Bird of*

Paradise, and parachute, the proposed algorithm achieves favorable results against existing methods [38, 40, 15]. Typically these methods aggregate superpixels from the whole image or consider temporal consistency within only two frames. The proposed method allows computing granularity inside the rough bounding box throughout several concussive frames to conduct MIL algorithm, thus temporal appearance change could be foreseen. Our method also favors video sequences where foreground and background are similar, for example *Penguin, Frog*, since temporal consistency is enhanced in multi-level.

Davis Dataset. Densely Annotated Video Segmentation Dataset was newly proposed in by Perazzi *et al*. [32] in CVPR 2016. It consists of 50 sequences with 3455 annotated frames, captured at 24*fps* and 1080p as well as 480p spacial resolution. All major challenges for video segmentation task could be found in this dataset, including background clutter, deformation, motion blur, scale-variation, camera shake, appearance change, etc. For each video frame, they provide a pixel-accurate, manually created segmentation ground truth in form of binary mask. Results of more than ten video segmentation algorithms are presented and compared.

Similar with Segtrack dataset, results are evaluated by Intersection-over-Union (IoU) for every frame. Table 2 lists accuracy for purposed algorithm and 6 existing methods. Overall, our algorithm achieves better results against other state-of-art methods. As seen in the table, the proposed method excels at treating challenging videos with complex appearance change (*swing, scooter-black,scooter-gray, rollerblade*), occlusions (*dog, bus*), motion blur (*dog, breakdance-flare*), etc. Figure 3 lists several examples of our segmentation results on this dataset.

In order to further analyse our multi-granularity model, we list in Table 3 tracking-generated bounding-box accuracy, MIL-superpixel level accuracy as well as final pixel-level accuracy for several sequences in Segtrack v2 and Davis dataset. Various conclusions can be drawn from this table. First, tracking accuracy greatly influences segmentation accuracy. Even though we expand bound boxes to sample negative bags and shrink them to sample positive ones, inaccurate bounding boxes could lead to large misclassification, leaving pixels around boundaries difficult to compensate mistakes. For sequence *Drift-straight,Stroller*, self-adapting bounding boxes fail to expand as much as foreground object, so the accuracy for this video falls respectively to 55.8% and 30%. However, using other tracking methods that favor great scale-variation could improve accuracy for this kind of video. Ideally speaking, when bounding boxes converge to tightly surrounding segmentation ground truth mask and MIL conditions are perfectly satisfied, the accuracy for same video is 91.2% and 88.3%, which, to some extent, proves the effectiveness of the proposed method and its promising performance with better tracking methods. Second, as shown in the table, our multi-granularity method can refine segmentation effectively. Even though in the proposed method, both spatial and temporal information are considered in MIL process, superpixels could still lack precision on borders, especially when foreground objects are not connected or have holes within. As on the table, pixel-level information help improve accuracy for listed videos by 20%. A vivid illustration is presented by Figure 4.

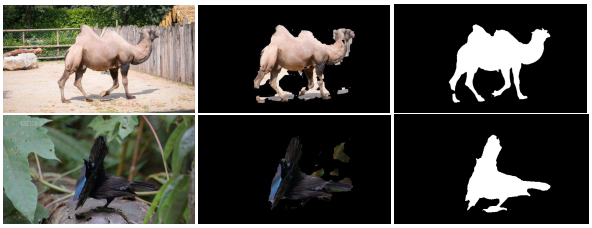

Figure 4. Demonstration results for our segmentation methods on Davis (top) and Segtrack (bottom) dataset. The first column is the original video frame. The second is the results of superpixel-parsing. Last column represents final refined results.

5. Conclusion

In this paper, we introduce a Multi-Granularity Analysis framework for video segmentation in a coarse-to-fine manner and prove that the segmentation problem can be easily solved. We apply the multiple instance learning in both the temporal and spatial domains to deal with the issues of temporal superpixel mismatch and noisy background within tracked bounding boxes, respectively. We show that our method performs favorably against state-of-art methods on popular datasets.

6. Acknowledgement

The work was supported by State Key Research and Development Program (2016YFB1001003). This work was partly supported by NSFC (61502301), China's Thousand Youth Talents Plan, National Natural Science Foundation of China (61521062), the 111 Project (B07022) and the Shanghai Key Laboratory of Digital Media Processing and Transmissions.

References

[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *TPAMI*, 34:2274–2282, 2012.

[2] P. Arbeláez, J. Pont-Tuset, J. T. Barron, F. Marques, and J. Malik. Multiscale combinatorial grouping. In *CVPR*, 2014.

[3] B. Babenko, M.-H. Yang, and S. Belongie. Robust object tracking with online multiple instance learning. *TPAMI*, 33:1619–1632, 2011.

[4] V. Badrinarayanan, F. Galasso, and R. Cipolla. Label propagation in video sequences. In *CVPR*, 2010.

[5] T. Brox and J. Malik. Large displacement optical flow: descriptor matching in variational motion estimation. *TPAMI*, 33:500–513, 2011.

[6] J. Chang, D. Wei, and J. W. Fisher. A video representation using temporal superpixels. In *CVPR*, 2013.

[7] M. Danelljan, F. Shahbaz Khan, M. Felsberg, and J. Van de Weijer. Adaptive color attributes for real-time visual tracking. In *CVPR*, 2014.

[8] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence*, 89:31–71, 1997.

[9] A. Faktor and M. Irani. Video segmentation by non-local consensus voting. In *BMVC*, 2014.

[10] Q. Fan, F. Zhong, D. Lischinski, D. Cohen-Or, and B. Chen. Jumpcut: non-successive mask transfer and interpolation for video cutout. *ACM Transactions on Graphics (TOG)*, 34:195, 2015.

[11] A. M. Ferman and A. M. Tekalp. Efficient filtering and clustering methods for temporal video segmentation and visual summarization. *Journal of Visual Communication and Image Representation*, 9:336–351, 1998.

[12] F. Galasso, R. Cipolla, and B. Schiele. Video segmentation with superpixels. In *ACCV*, 2012.

[13] F. Galasso, M. Keuper, T. Brox, and B. Schiele. Spectral graph reduction for efficient image and streaming video segmentation. In *CVPR*, 2014.

[14] F. Galasso, N. Shankar Nagaraja, T. Jimenez Cardenas, T. Brox, and B. Schiele. A unified video segmentation benchmark: Annotation, metrics and analysis. In *ICCV*, 2013.

[15] M. Grundmann, V. Kwatra, M. Han, and I. Essa. Efficient hierarchical graph-based video segmentation. In *CVPR*, 2010.

[16] S. D. Jain and K. Grauman. Supervoxel-consistent foreground propagation in video. In *ECCV*, 2014.

[17] W.-D. Jang and C.-S. Kim. Streaming video segmentation via short-term hierarchical segmentation and frame-by-frame markov random field optimization. In *ECCV*, 2016.

[18] J. D. Keeler, D. E. Rumelhart, and W.-K. Leow. *Integrated Segmentation and Recognition of Hand-Printed Numerals*. Microelectronics and Computer Technology Corporation, 1991.

[19] A. Khoreva, F. Galasso, M. Hein, and B. Schiele. Classifier based graph construction for video segmentation. In *CVPR*, 2015.

[20] Y. J. Lee, J. Kim, and K. Grauman. Key-segments for video object segmentation. In *ICCV*, 2011.

[21] V. Lempitsky, P. Kohli, C. Rother, and T. Sharp. Image segmentation with a bounding box prior. In *ICCV*, 2009.

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

[23] T.-C. Lin, M.-C. Yang, C.-Y. Tsai, and Y.-C. F. Wang. Query-adaptive multiple instance learning for video instance retrieval. volume 24, pages 1330–1340, 2015.

[24] C. Ma, J.-B. Huang, X. Yang, and M.-H. Yang. Hierarchical convolutional features for visual tracking. In *ICCV*, 2015.

[25] T. Ma and L. J. Latecki. Maximum weight cliques with mutex constraints for video object segmentation. In *CVPR*, 2012.

[26] N. Märki, F. Perazzi, O. Wang, and A. Sorkine-Hornung. Bilateral space video segmentation. In *CVPR*, 2016.

[27] O. Maron and T. Lozano-Pérez. A framework for multiple-instance learning. *NIPS*, 1998.

[28] H. Nam and B. Han. Learning multi-domain convolutional neural networks for visual tracking. *arXiv preprint arXiv:1510.07945*, 2015.

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

[30] P. Ochs, J. Malik, and T. Brox. Segmentation of moving objects by long term video analysis. *TPAMI*, 36:1187–1200, 2014.

[31] A. Papazoglou and V. Ferrari. Fast object segmentation in unconstrained video. In *ICCV*, 2013.

[32] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. V. Gool, M. Gross, and A. Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *CVPR*, 2016.

[33] F. Perazzi, O. Wang, M. Gross, and A. Sorkine-Hornung. Fully connected object proposals for video segmentation. In *ICCV*, 2015.

[34] R. Sanz-Requena, D. Moratal, D. R. García-Sánchez, V. Bodí, J. J. Rieta, and J. M. Sanchis. Automatic segmentation and 3d reconstruction of intravascular ultrasound images for a fast preliminar evaluation of vessel pathologies. *Computerized Medical Imaging and Graphics*, 31:71–80, 2007.

[35] R. E. Schapire and Y. Singer. Improved boosting algorithms using confidence-rated predictions. *Machine learning*, 37:297–336, 1999.

[36] G. Seguin, P. Bojanowski, R. Lajugie, and I. Laptev. Instance-level video segmentation from object tracks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[37] J. Shi and J. Malik. Motion segmentation and tracking using normalized cuts. In *ICCV*, 1998.

[38] Y.-H. Tsai, M.-H. Yang, and M. J. Black. Video segmentation via object flow. In *CVPR*, 2016.

[39] S. Vijayanarasimhan and K. Grauman. Active frame selection for label propagation in videos. In *ECCV*, 2012.

[40] S. Wang, H. Lu, F. Yang, and M.-H. Yang. Superpixel tracking. In *ICCV*, 2011.

[41] L. Wen, D. Du, Z. Lei, S. Z. Li, and M.-H. Yang. Jots: Joint online tracking and segmentation. In *CVPR*, 2015.

[42] J. Wu, Y. Zhao, J.-Y. Zhu, S. Luo, and Z. Tu. Milcut: A sweeping line multiple instance learning paradigm for interactive image segmentation. In *CVPR*, 2014.

[43] Y. Wu, J. Lim, and M. Yang. Online object tracking: A benchmark. In *CVPR*, 2013.

[44] C. Xu, C. Xiong, and J. J. Corso. Streaming hierarchical video segmentation. In *ECCV*, 2012.

[45] Y. Xu, J. Y. Zhu, I. C. Chang, M. Lai, and Z. Tu. Weakly supervised histopathology cancer image segmentation and classification. *Medical Image Analysis*, 18(3):591, 2014.

[46] S. Yi and V. Pavlovic. Multi-cue structure preserving mrf for unconstrained video segmentation. In *ICCV*, 2015.

[47] C. Zhang, J. C. Platt, and P. A. Viola. Multiple instance boosting for object detection. In *NIPS*, 2005.

[48] K. Zhang and H. Song. Real-time visual tracking via online weighted multiple instance learning. *Pattern Recognition*, 46:397–411, 2013.

[49] Q. Zhang and S. A. Goldman. Em-dd: An improved multiple-instance learning technique. In *NIPS*, 2001.