Weakly Supervised Multiclass Video Segmentation

Xiao Liu[1], Dacheng Tao[2], Mingli Song[1], Ying Ruan[1], Chun Chen[1] and Jiajun Bu[1]
[1]*Zhejiang Provincial Key Laboratory of Service Robot, Zhejiang University, China*
{ender_liux, brooksong, yingruan, chenc, bjj}@zju.edu.cn

[2]*Centre for Quantum Computation and Intelligent Systems*
Faculty of Engineering and Information Technology
University of Technology, Sydney, Australia
dacheng.tao@uts.edu.au

Abstract

The desire of enabling computers to learn semantic concepts from large quantities of Internet videos has motivated increasing interests on semantic video understanding, while video segmentation is important yet challenging for understanding videos. The main difficulty of video segmentation arises from the burden of labeling training samples, making the problem largely unsolved. In this paper, we present a novel nearest neighbor-based label transfer scheme for weakly supervised video segmentation. Whereas previous weakly supervised video segmentation methods have been limited to the two-class case, our proposed scheme focuses on more challenging multiclass video segmentation, which finds a semantically meaningful label for every pixel in a video. Our scheme enjoys several favorable properties when compared with conventional methods. First, a weakly supervised hashing procedure is carried out to handle both metric and semantic similarity. Second, the proposed nearest neighbor-based label transfer algorithm effectively avoids overfitting caused by weakly supervised data. Third, a multi-video graph model is built to encourage smoothness between regions that are spatiotemporally adjacent and similar in appearance. We demonstrate the effectiveness of the proposed scheme by comparing it with several other state-of-the-art weakly supervised segmentation methods on one new Wild8 dataset and two other publicly available datasets.

1. Introduction

Video segmentation, the problem of assigning labels for pixels in a video sequence, is an important computer vision task with applications in areas such as advertisement recommendation, activity recognition, video summarization and target retrieval. Despite its significance, the problem is largely under-addressed due to the heavy burden of labeling training samples. Learning a fully supervised segmentation model requires the labeling of every pixel in every frame of the training videos, which is very time-consuming and labor-intensive.

To reduce the labeling burden of the fully supervised model, semi-supervised video segmentation methods are proposed [6] whereby only part of the training frames are required to be labeled. Nevertheless, although the labeling tasks can be easier in semi-supervised methods than that in fully supervised ones, it is still very difficult to obtain enough labeled training data because of the difficulty of pixel-level labeling. To avoid pixel-level labeling, weakly supervised video segmentation methods are proposed [14, 27] in which semantic labels are assigned to videos but the labels are not spatially or temporally localized within the videos. Video-level labeling is much easier to do than pixel-level labeling but the absence of localization information leads to training data ambiguity and limits the application of weakly supervised methods. Note that the data ambiguity of weakly supervised video segmentation is more serious than that of weakly supervised image segmentation since using more pixels leads to more possible labelings and is easier to cause training overfitting.

Thus, most margin maximization-based weakly supervised image segmentation methods [30, 31, 32, 36] are not suitable for video segmentation; and existing weakly supervised video segmentation methods [14, 27] only achieve convincing results for the task of two-class video segmentation, i.e. separating the foreground region from the background, but they are not suitable for multiclass video segmentation. Compared with the two-class case, multiclass video segmentation faces the following challenges: (1) easy to mix up multiple classes in the metric space, (2) easy to cause overfitting in multiclass label prediction and (3) difficult to conduct appearance-based multiclass classification.

To conquer these challenges, we propose a nearest neighbor-based label transfer scheme for weakly supervised multiclass video segmentation. According to [35], videos are first segmented into supervoxels, from which color, pattern, texture, and motion features are extracted. We then apply a weakly supervised hashing procedure to transform the features to compact binary codes, such that the semantic similarity of two supervoxels can be efficiently calculated by the Hamming distance between their binary codes. We develop the nearest neighbor-based label transfer algorithm which works under the following intuition: if two supervoxels from different videos are similar to each other, and the two videos share some video-level labels, it is reasonable to assume that both supervoxels share the same labels. Considering all the pairwise relationships between such supervoxels, the categorical probability of a supervoxel can be estimated by transferring the video-level labels from its neighbors. We show that a key improvement of the proposed algorithm over conventional algorithms is that it effectively avoids overfitting caused by weakly supervised data. Using the appearance-based categorical probability as the unary energy, we build a multi-video graph model to encourage smoothness between spatiotemporally adjacent supervoxels in the same video and supervoxels of similar appearance across the videos. Following the above procedure, the video-level labels are transformed to the pixel level. By using standard supervised methods, we can further utilize the resultant pixel-level labeled videos to segment new videos.

2. Related Work

2.1. Video Segmentation

Impressive progress has been reported on unsupervised video segmentation with methods ranging from hierarchical graph model [12], streaming graph model [35], region tracking [3], interactive matting [1] hypergraph cut [15] and multiple hypothesis tracking [4]. These methods successfully segment videos into consistent regions, namely supervoxels. However, due to the absence of supervision information, these unsupervised segmentation methods cannot associate the supervoxels with semantic meanings, so their application is limited to low-level video processing rather than high-level semantic understanding.

By contrast, supervised classifiers such as random forest [2] show promising results for semantic video segmentation [5, 24, 22], although fully supervised training classifiers require copious quantities of labeled video data, which are extremely arduous to obtain by hand labeling. To reduce the burden of labeling, semi-supervised methods are introduced for video segmentation [6, 29]. Given a section of hand-labeled frames from a video sequence, semi-supervised methods propagate labels throughout the rest of the video sequence. Although semi-supervised methods require less labeling work than supervised methods, it is still very difficult to obtain enough labeled training data because of the difficulty of pixel-level labeling. To avoid pixel-level labeling, weakly supervised video segmentation is proposed [14, 27] whereby semantic labels are associated with training videos but the labels are not spatially or temporally localized within the video. Through discriminative learning [14] and concept ranking [27], both methods separate the foreground from the background but do not apply for multiclass video segmentation.

2.2. Weakly Supervised Image Segmentation

Weakly supervised image segmentation has been profoundly investigated, such as [30, 31, 32, 37, 36, 11, 28]. Vasconcelos et al. [28] transformed the weakly supervised segmentation problem to an annotation problem by applying unsupervised normalized cut segmentation [23] before assigning labels. The method is easy to implement but only feasible for simple tasks. Vezhnevets and Buhmann [30] exploited a weakly supervised random forest for image segmentation. It has been shown that directly maximizing the classification margin for weakly supervised data leads to learning overfitting, and in [30] additional labeled data are used to regularize the random forest construction to prevent this. Consequently, [30] still needs a large quantity of pixel-level labeled data for training. Vezhnevets et al. [31] extended [30] by introducing a multi-image model, in which the smoothness between adjacent and similar superpixels is encouraged. Since the unary energy of [31] depends on the weakly supervised random forest method proposed in [30], however, additional hand labeled data are also required to regularize the model.

Vezhnevets et al. [32] proposed a weakly supervised method for both metric learning and image segmentation. Since it is based on alternating optimization, the results are sensitive to model initialization and training overfitting may be caused by bad initialization. The proposed scheme in this paper also focuses on the task of metric learning and segmentation. Compared to the above-mentioned methods, it does not depend on pixel-level data or alternating optimization. Thus, it is more efficient and more robust to ambiguous data and noise than [32].

3. Weakly Supervised Multiclass Video Segmentation

Weakly labeled videos are first segmented into spatiotemporal supervoxels, which are represented as high dimensional points in the feature space, and weakly supervised hashing is subsequently carried out for metric learning. The proposed nearest neighbor-based label transfer algorithm is then used to estimate categorical probabilities,

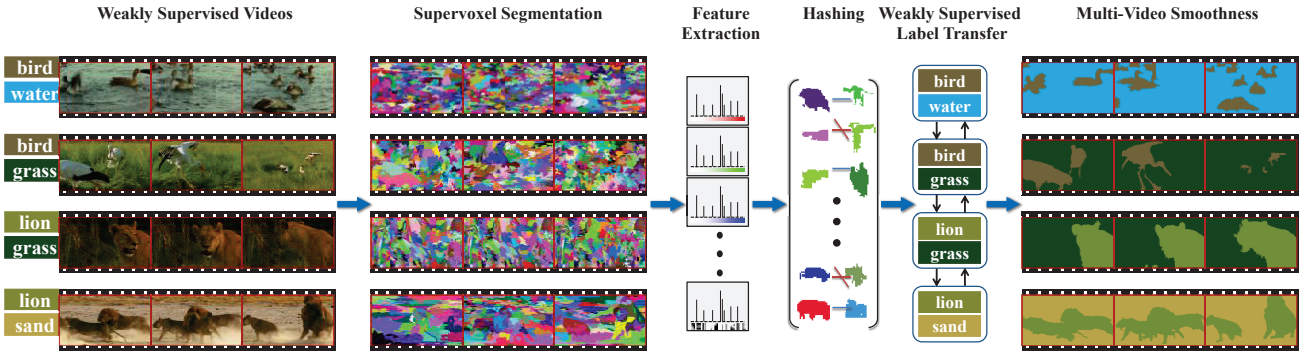


Figure 1. Flowchart of the proposed weakly supervised multiclass video segmentation scheme. Weakly labeled videos are first segmented into spatiotemporal supervoxels, which are represented as high dimensional points in the feature space, and weakly supervised hashing is subsequently carried out for metric learning. The proposed nearest neighbor-based label transfer algorithm is then used to estimate categorical probabilities, and the final pixel-level labels are decided by the multi-video model.

and the final pixel-level labels are decided by the multi-video model. We omit the step of segmenting new videos, since this can be regarded as a standard supervised learning problem. Figure 1 shows the flow of the proposed scheme.

The given dataset can be represented as $\tau = \{B^j = (\{x_i^j\}_{i=1}^{N_j}, Y^j)\}_{j=1}^{N}$ where supervoxel x_i^j comes from video B^j. Each video B^j has a label set Y^j, which is a subset of the full label set $\left(Y^j \subset Y = \{1, \ldots, C\}\right)$ corresponding to semantic concepts. Every supervoxel x_i^j is associated with a latent label $y_i^j \in Y^j$. The bag label set Y^j contains the labels of all supervoxels in that bag $(Y^j = \cup y_i^j)$. N is the number of videos and N_j is the number of supervoxels in the jth video. We use f_i^j to indicate the feature of x_i^j. For ease of notation, we also sequence the supervoxels from 1 to L, and use a matrix $X = [x_1 \ldots x_L]$ to indicate the input supervoxels.

3.1. Weakly Supervised Hashing

A key factor of a segmentation algorithm is the metric which properly measures dissimilarity between supervoxels. The L_2 distance based on human-defined features, e.g. SIFT, color histogram, LBP and optical flow, is the most straightforward metric for measuring the dissimilarity between supervoxels. In the context of weakly supervised video segmentation however, the L_2 distance on human-defined visual features is not optimal due to the ignorance of two valuable observations: spatiotemporally adjacent cues and the high-level semantic information.

Regarding smoothness, if two supervoxels are spatiotemporally adjacent and are neighbors in the feature space, the two supervoxels belong to the same semantic concept. By contrast, if two videos do not share any semantic labels, none of the supervoxel pairs across the two videos belong to the same semantic concept.

Combining the two observations, two types of pairwise relationship sets can be built up using the weakly supervised video data: the link relationship set (M) and the non-link relationship set (V). Specifically, for a video B^j, a pair of supervoxels $(x_i^j, x_k^j) \in M$ if x_i^j and x_k^j are spatiotemporally adjacent and $||f_i^j - f_k^j||_2 < \epsilon$. For a pair of videos B^{j_1} and B^{j_2}, the supervoxel pairs $(x_i^{j_1}, x_k^{j_2})_{i=1\ldots N_{j_1}, k=1\ldots N_{j_2}} \in V$ if $Y^{j_1} \cap Y^{j_2} = \emptyset$. The above two relationship sets are represented by defining a matrix $S \in R^{L \times L}$ that incorporates the pairwise information

$$S_{ij} = \begin{cases} 1 & : \quad (x_i, x_j) \in M \\ -1 & : \quad (x_i, x_j) \in V \\ 0 & : \quad \text{otherwise.} \end{cases} \tag{1}$$

Regarding the running speed in our implementation, we sample a subset of supervoxels to construct S instead of using all the supervoxels. Through constructing S, we transform the video-level semantic labels to a supervoxel-level relationship matrix that is used for metric learning. We follow the spirit of semi-supervised hashing [33, 17] by generating binary codes for metric learning, because it efficiently handles both metric and semantic similarity. A sequence of K hashing functions $H = [h_1, \ldots, h_K]$ are learned and each hashing function maps the original features to either 1 or -1. $H(X) \in R^{K \times L}$ is the mapped binary codes of X. An objective function that measures the conformity between the binary codes and S is maximized, while the balance and independence of the binary codes are guaranteed

$$\begin{aligned} H^* &= \underset{H}{\operatorname{argmax}} \operatorname{tr}\left\{H(X) S H(X)^\top\right\}, \\ s.t. \quad & h_k(x) = 0, \quad k = 1 \ldots K, \\ & H(X) H(X)^{\mathrm{T}} = LI. \end{aligned} \tag{2}$$

The objective function is difficult to solve, but its relaxed form can be efficiently optimized by the spectral method.

After the weakly supervised hashing procedure, a supervoxel x_i^j is represented as a sequence of binary codes b_i^j, and the dissimilarity between a pair of supervoxels $x_{i_1}^{j_1}$ and $x_{i_2}^{j_2}$ is measured by the Hamming distance

$$d_H(x_{i_1}^{j_1}, x_{i_2}^{j_2}) = \frac{1}{4}||b_{i_1}^{j_1} - b_{i_2}^{j_2}||_2. \tag{3}$$

3.2. Nearest Neighbor-based Label Transfer

All supervised learning methods rely on the smoothness assumption [7]: if two points in the feature space are close, the corresponding labels should also be similar. Considering a supervoxel-label pair (x_i^j, y_i^j), if we have another point $x_{i*}^{j^*}$ that is close to x_i^j, we may expect to predict y_i^j by $y_{i*}^{j^*}$. In the weakly supervised context, we rarely know the category label of any training sample, so the naïve nearest neighbor-based method does not work.

However, if two supervoxels from different videos are similar to each other, and the two videos share some video-level labels, it is reasonable to conclude from the smoothness assumption that both supervoxels share the same labels. If x_i^j is close to $x_{i*}^{j^*}$ and $Y^j \cap Y^{j^*} \neq \emptyset$, we transfer the video-level labels $Y^j \cap Y^{j^*}$ to the supervoxel labels y_i^j, and the categorical probability of a given supervoxel x_i^j is

$$p(y_i^j|x_i^j) = \frac{1}{|\mathbf{N}_{ap}(x_i^j)|} \sum_{x_{i*}^{j^*} \in \mathbf{N}_{ap}(x_i^j)} \frac{\delta\left(y_i^j \in \left(Y^j \cap Y^{j^*}\right)\right)}{|Y^j \cap Y^{j^*}|}, \tag{4}$$

where $\mathbf{N}_{ap}(\cdot)$ denotes the appearance-based neighbor set and $\delta(\cdot)$ is the indicator function. Each neighbor of x_i^j has the same weight $(1/|\mathbf{N}_{ap}(x_i^j)|)$ for estimation and each shared label contributes equally. If most neighbors of x_i^j have the same video-level labels, y_i^j will have a high probability in that category.

In this paper, instead of searching for a constant number of nearest neighbors, a supervoxel $x_{i*}^{j^*}$ is regarded as the neighbor of x_i^j if it satisfies three conditions: (1) it is sufficiently similar to x_i^j,

$$d_H(x_i^j, x_{i*}^{j^*}) < \sigma, \tag{5}$$

(2) of all the supervoxels in video B^j, x_i^j is the most similar supervoxel to $x_{i*}^{j^*}$

$$x_i^j = \underset{x_k^j, k=1\ldots N_j}{\operatorname{argmin}} \; d_H(x_i^j, x_{i*}^{j^*}), \tag{6}$$

and (3) B^j and B^{j^*} share some labels

$$Y^j \cap Y^{j^*} \neq \emptyset. \tag{7}$$

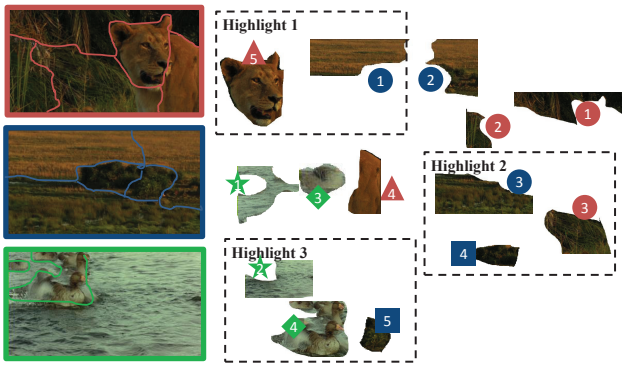


Figure 2. Intuition behind the three conditions of finding neighbors for nearest neighbor-based label transfer. The colors indicate different videos, and the various shapes indicate supervoxels taken from different categories. Square, circle, triangle, star and rhombus indicate *tree*, *grass*, *lion*, *water* and *bird*, respectively. Refer to the text for detailed explanations.

Figure 2 details the intuition behind the three conditions of finding neighbors for nearest neighbor-based label transfer. The Highlight 1 area shows the requirement of the first condition of finding neighbors: the red triangle 5 is far from all the blue points, so there is no indication of whether or not it is circle.

Since appearances of some categories (e.g. *tree* and *grass*) are relatively similar, the distance between two points from such categories can be smaller than the threshold. In this case, using the first condition alone will lead to a wrong determination of neighbors. The Highlight 2 area in Figure 2 shows this situation: although the blue square 4 is close to the red circle 3, it is from the square category, not the circle category. Although two points from different categories may be close, they are rarely the closest points. By contrast, since points from the same category tend to fall near one another, they are likely to be the closest points from the same category. By this analysis, our second condition suppresses all the points that are not the closest such that the algorithm is robust to noisy points.

The Highlight 3 area in Figure 2 illustrates the third condition for finding neighbors: the green and blue points share no common labels, so although the green points are near the blue points, they are not from the same category and are not regarded as neighbors for label transfer.

It should be noted that conventional weakly supervised segmentation methods [14, 27] use all the training data to make predictions, e.g. when labeling the *bird* areas in the 1st row of Figure 1, all the other videos are used to learn the appearance model of *bird*. As the videos from the 3rd and 4th rows do not have *bird*, supervoxels in these videos are regarded as negative samples. However, since appearances of the categories (*lion*, *grass*, *sand*) of the videos are

very different from the appearance of *water*, we cannot expect the resultant classifier to correctly distinguish *bird* and *water*. Hence, The training error may be small (it can correctly classify *bird*+*water* against *lion*+*grass*+*sand*), but the testing error can be very large (it fails to classify *bird* against *water*). In other words, using such videos for training causes overfitting in multiclass video segmentation. In our proposed algorithm, by contrast, videos without common labels do not help each other in prediction, and overfitting can be efficiently avoided.

3.3. Multi-Video Graph Model

Given the input weakly supervised videos $\tau = \{B^j = (\{x_i^j\}_{i=1}^{N_j}, Y^j)\}_{j=1}^{N}$, the categorical posterior is

$$\begin{aligned} & p\left(\{y_i^j\}|\{x_i^j\},\{Y^j\}\right) \\ & \propto p\left(\{y_i^j\}|\{x_i^j\}\right) p\left(\{Y^j\}|\{y_i^j\}\right), \end{aligned} \tag{8}$$

which has two factors. The first factor

$$\begin{aligned} p\left(\{y_i^j\}|\{x_i^j\}\right) &= p\left(\{y_i^j\}\right) \prod\nolimits_{j,i} p\left(x_i^j|y_i^j\right) \\ &\propto p\left(\{y_i^j\}\right) \prod\nolimits_{j,i} p\left(y_i^j|x_i^j\right) \end{aligned} \tag{9}$$

is a posterior that can be formulated by a Markov Random Fields (MRF) model. Based on the Hammersley-Clifford theorem [13], the configuration factor $p(\{y_i^j\})$ is a Gibbs distribution, and then (9) can be simplified as

$$\begin{aligned} & p\left(\{y_i^j\}|\{x_i^j\}\right) \\ & \propto \exp\left(\sum\nolimits_{j,i} T \log p\left(y_i^j|x_i^j\right) - U\left(\{y_i^j\}\right)\right) \\ & = \exp\left(-E\left(\{y_i^j\}\right)\right), \end{aligned} \tag{10}$$

where T is the temperature parameter, $U(\{y_i^j\})$ is the prior potential energy, and $E(\{y_i^j\})$ is the posterior energy. In this paper, two types of configuration priors are encouraged. The first prior is the spatiotemporal smoothness, which regards all the spatiotemporally adjacent supervoxels of a supervoxel x_i^j as its spatiotemporal-neighbor $\mathbf{N}_{st}(x_i^j)$. The second prior is the smoothness between similar supervoxels, and we use $\mathbf{N}_{ap}(x_i^j)$ defined in Section 3.3 as the appearance-neighbor set of x_i^j. The pairwise smoothness cost has the form of a contrast sensitive Potts model [20]

$$v\left(y_{i_1}^{j_1}, y_{i_2}^{j_2}\right) = \left(K - d_H\left(x_{i_1}^{j_1}, x_{i_2}^{j_2}\right)\right) \delta\left(y_{i_1}^{j_1} \neq y_{i_2}^{j_2}\right). \tag{11}$$

and the posterior energy can be written as

$$\begin{aligned} E\left(\{y_i^j\}\right) = & \sum_{j=1\ldots N, i=1\ldots N_j} \Big(-T \log p\left(y_i^j|x_i^j\right) \\ & + \sum\nolimits_{x_{i*}^{j*} \in \left(\mathbf{N}_{st}\left(x_i^j\right) \cup \mathbf{N}_{ap}\left(x_i^j\right)\right)} v\left(y_i^j, y_{i*}^{j*}\right)\Big). \end{aligned} \tag{12}$$

The second factor of (8) is

$$p\left(\{Y^j\}|\{y_i^j\}\right) = \begin{cases} 1 & : \quad \forall j, \quad Y^j = \cup\{y_k^j\}_k \\ 0 & : \quad \text{otherwise}, \end{cases} \tag{13}$$

which guarantees the weakly supervised restrictions. Since assigning labels that do not appear in a video causes infinity posterior energy, the only remaining concern is that all the video-level labels should be assigned to supervoxels. This restriction can be guaranteed by simply choosing a supervoxel with the largest categorical posterior for each video-level label and fixing the label.

Finally, we inference the posterior energy by the coarse-to-fine supervoel trees [16] and the alpha-enpension [26] algorithm is used for optimization, such that the optimized supervoxel-level labels maximize the posterior.

4. Experiments

We validate the proposed scheme on one new Wild8 dataset, the YTO dataset and the SUNY 24-class Dataset.

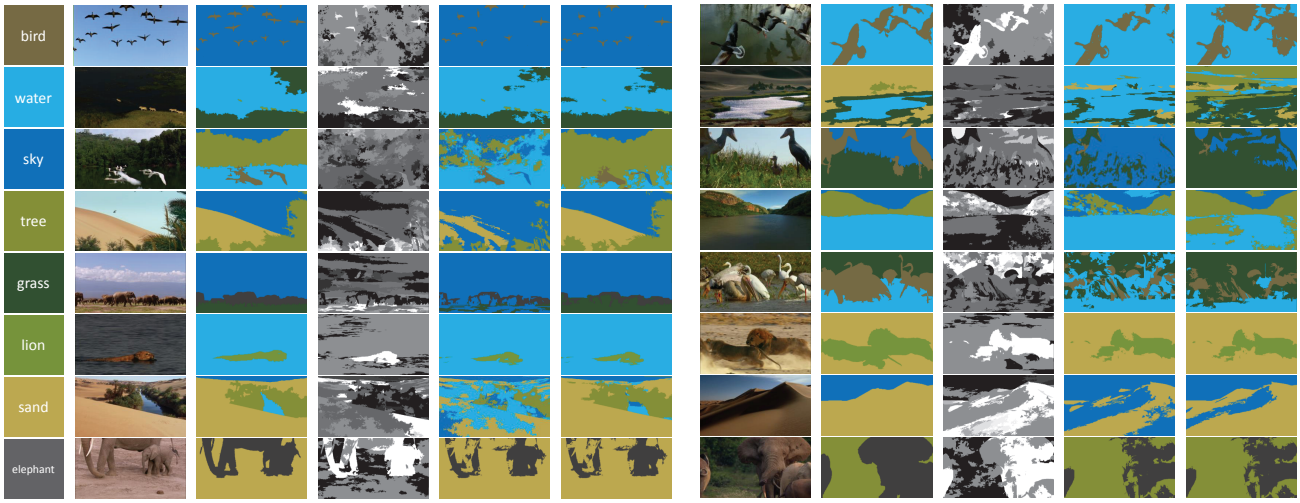
4.1. Wild8 Dataset

We construct a "Wild8" dataset to quantitatively evaluate the proposed scheme due to the lack of a multiclass video segmentation dataset with semantical pixel-level ground-truth. Our dataset consists of 100 sequences of weakly supervised videos from 3 documentary series[1] of which 33 sequences are manually labeled with pixel-level ground-truth for evaluation. The dataset covers 8 categories (*bird*, *lion*, *elephant*, *sky*, *tree*, *grass*, *sand*, and *water*) and all the sequences are associated with multiple categories. Each sequence has a length of three seconds and is sampled at 10Hz. All the frames are resized to the same dimensions (640×480) for ease of processing.

We utilize the streamGBH algorithm [35] implemented by the publicly available LIBSVX library [34] for supervoxel segmentation. To balance the preservation of the object boundary and spatiotemporal uniformity, we choose the 15th hierarchy as the finest level in building the supervoxel trees and set the other parameters as their default values.

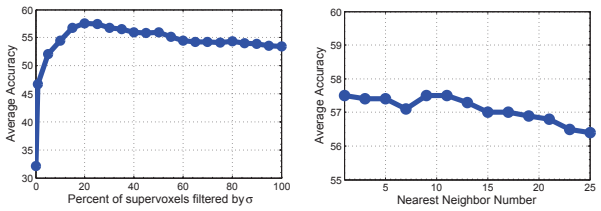
We then represent the supervoxels using the following set of features: RGB color histogram quantized into 48 bins, local binary patterns quantized into 59 bins [19], optical flow histogram quantized into 5 bins [8], gradient histogram quantized into 9 bins [18], and densely computed heatmaps. The distance threshold ϵ for hashing is set to allow the top 30% of similar supervoxels to be neighbors. The number of hashing functions K is set to 128 because our experiments show that using more hashing functions improves results slightly but increases the computational time. The temperature parameter T is empirically set to 10^5.

[1] "Winged Migration" (Jacques Perrin et al., 2001), "The Last Lions" (Dereck Joubert, 2011), and "Africa" (David Attenborough, 2013).



(a) (b)

Figure 4. Samples of our segmentation results on the Wild8 dataset. We use different colors to indicate the categories in the first column, and list sample frames of weakly supervised videos in the second column, while the ground-truths are given in the third column. The categorical probabilities (a lighter area indicates higher probability) and segmentation results before and after smoothing are given in the fourth to sixth columns. We show successful segmentation results in (a) and partially successful segmentation results in (b).



(a) (b)

Figure 3. Average per class accuracy over different values of σ (a) and different nearest neighbor numbers (b).

We test various values of the filtering threshold σ in (5). A larger σ allows the algorithm to use more (but noisy) clues, while a smaller σ tends to use the most promising clues. It emerges that the best choice depends on the size of the database. For the Wild8 dataset, we set σ to allow the top 20% similar supervoxels to empirically be neighbors (Figure 3a). We try different nearest neighbor numbers in (6) instead of only using the most similar neighbors. The experiment shows that a large number of nearest neighbors (e.g. 10+) will cause a slight drop in performance. If the appearance of some categories is similar, the distance between two supervoxels from such categories will be small, and a large number of nearest neighbors may thus lead to the incorrect determination of neighbors.

We compare the proposed scheme with the state-of-the-art weakly supervised video segmentation methods, including CRANE [27], MIN [25], SVM [14] and MIL [31]. For CRANE and MIN, we first collect positive (with noise) and negative samples for each category, and then apply concept ranking. For SVM, we use LIBLINEAR [10] to train a one-against-all classifier for each category. For MIL, we use randomly initialized AmmMIL-RF+PPinv [30] with multi-video smoothness. All the methods take the same supervoxels and features as input. For fair comparison, our scheme without hashing or smoothness are also tested. We summarize the segmentation accuracies of the 8 categories, the average per category accuracy (ave_acc), total per pixel accuracy (tot_acc), and mean average precision (mAP) in Table 1. Note that the mAP must be calculated before smoothing.

Our scheme outperforms all existing weakly supervised video segmentation methods, especially in the difficult tasks of segmenting *lion* and *elephant*. MIL [31] achieves the highest accuracy on the *water* and *grass* categories since the margin maximization algorithm tends to perform well on larger categories, at the expense of the poor performance on small categories, such as *bird*, *lion* and *elephant*. We can see that the weakly supervised hashing brings an improvement of 3.3% on the ave_acc while the multi-video smoothing brings an improvement of 0.7%.

Samples of our segmentation results on the Wild8 dataset are shown in Figure 4. In most cases, our algorithm gives accurate categorical probabilities, e.g the 1st, and 4th-8th rows in Figure 4(a). The 3rd row in Figure 4(a) shows the power of the smoothness, while the 4th row in Figure 4(b) gives a negative example. In general, the multi-video smoothness slightly improves segmentation performance but more importantly makes the visual appearance of the segmentation look good.

Method	bird	water	sky	tree	grass	lion	sand	elephant	ave_acc	tot_acc	mAP
CRANE [27]	47.8	76.5	89.5	42.8	73.7	19.3	43.2	16.8	51.2	62.4	43.9
MIN [25]	48.1	75.2	87.2	36.7	74.1	15.4	43.3	13.2	49.2	60.8	42.1
SVM [14]	42.5	74.5	86.9	45.5	74.0	16.6	42.1	12.3	49.3	61.2	41.0
MIL [31]	31.5	**79.3**	85.4	41.1	**78.3**	2.1	55.2	5.5	47.3	62.9	41.8
Ours without hashing	50.9	73.7	90.4	45.9	74.4	17.7	56.7	24.1	54.2	65.1	47.5
Ours without smoothness	**53.5**	77.1	92.6	50.1	75.6	20.9	58.8	**28.3**	57.1	67.7	**52.4**
Ours	53.0	77.3	**93.8**	**50.1**	76.5	**21.3**	**60.1**	28.1	**57.5**	**68.4**	-

Table 1. Results of the Wild8 dataset. We measure the segmentation accuracy of each class, average per category accuracy (ave_acc), total per pixel accuracy (tot_acc) and mean average precision (mAP). When compared with the state-of-the-art, our scheme outperforms all existing weakly supervised video segmentation methods.

Method	CRANE	MIN	SVM	MIL	Ours
mAP	42.5	37.7	31.2	38.5	46.1

Table 2. mAP scores of different weakly supervised segmentation methods on the YTO dataset.

4.2. YTO Dataset

The YouTube-Objects (YTO) dataset [21] is composed of videos from YouTube over 10 object classes (*aeroplane*, *bird*, *boat*, *car*, *cat*, *cow*, *dog*, *horse*, *motorbike* and *train*). There are up to 24 videos for each class, of which the duration varies from 30 seconds to 3 minutes. The videos have been weakly annotated by hand. We test our scheme on this dataset to validate its performance for large scale data.

It should be mentioned that although the YTO dataset contains videos from multiple classes, each video belongs to only one class and only objects of the relevant class can be presented in the video. We therefore denote the task on this dataset as a two-class segmentation rather than a multiclass segmentation since different classes do not interact with each other.

To apply our multiclass algorithm on the YTO dataset, we assign a common latent "background" label to all the videos so that all the videos share one label and have the same weight for label transfer. In this case, our nearest neighbor-based label transfer scheme is very similar to the CRANE algorithm which can be deemed as a special case of our scheme for two-class segmentation. We omit the multi-video smoothing step in this dataset since it takes too much time and memory to apply global optimization in such a big graph.

We report the mAP scores for different weakly supervised video segmentation methods in Table 2. All the methods use the same supervoxels and features and are tested on the same labeled videos provided by [27]. Our scheme achieves the highest mAP score among state-of-the-art methods and its improvement over CRANE is the result of the additional metric learning procedure and the more carefully designed nearest neighbor set construction.

4.3. SUNY Dataset

The SUNY 24-class dataset [9] is a collection of general-purpose sequences from Xigh.org. The dataset contains 8 videos with pixel-level labels. Compared with the Wild8 dataset, the SUNY dataset is more challenging because: (1) the number of its sequences is smaller, (2) each sequence has more labels, and (3) some labels only appear in one sequence. We use all 8 sequences and their video-level labels to inference the pixel-level labels. The proposed scheme achieves 14.1% ave_acc on this dataset and the CRANE algorithm achieves 13.8% ave_acc on this dataset.

5. Conclusion

This paper introduced a new nearest neighbor-based label transfer scheme for the challenging task of weakly supervised multiclass video segmentation. In this scheme, the weakly supervised hashing procedure for metric learning handles both metric and semantic similarity; the nearest neighbor-based label transfer algorithm suppresses the overfitting problem; and the multi-video graph model encourages smoothness between supervoxels that are spatiotemporally adjacent and similar in appearance. The success of the proposed scheme boosts video segmentation for practical applications including video advertisement recommendation, object tracking and recognition, and semantic level video summarization.

ACKNOWLEDGEMENTS

This work was supported in part by the National Natural Science Foundation of China (61170142), the National Key Technology R&D Program under Grant (2011BAG05B04), the Program of International S&T Cooperation (2013DFG12840), and Australian Research Council Projects (FT-130101457 and DP-140102164). M. Song is the corresponding author.

References

[1] X. Bai and G. Sapiro. Geodesic matting: A framework for fast interactive image and video segmentation and matting. *IJCV*, 82:113–132, 2009. 4322

[2] L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001. 4322

[3] W. Brendel and S. Todorovic. Video object segmentation by tracking regions. *Proc. ICCV*, pages 833–840, 2009. 4322

[4] W. Brendel and S. Todorovic. Multiple hypothesis video segmentation from superpixel flows. *Proc. ECCV*, pages 268–281, 2010. 4322

[5] G. Brostow, J.Shotton, J. Fauqueur, and R. Cipolla. Segmentation and recognition using structure from motion point clouds. *Proc. ECCV*, pages 44–57, 2008. 4322

[6] I. Budvytis, V. Badrinarayanan, and R. Cipolla. Label propagation in complex video sequences using semi-supervised learning. *Proc. BMVC*, pages 27.1–27.12, 2010. 4321, 4322

[7] O. Chapelle, B. Schölkopf, and A. Zien. *Semi-supervised Learning*. The MIT Press, Cambridge, MA, 2006. 4324

[8] R. Chaudhry, A. Ravichandran, G. Hager, and R. Vidal. Histograms of oriented optical flow and Binet-Cauchy kernels on nonlinear dynamical systems for the recognition of human action. *Proc. CVPR*, pages 1932–1939, 2009. 4325

[9] A. Chen and J. Corso. Propagating multi-class pixel labels throughout video frames. *Western New York Image Processing Workshop(WNYIPW)*, 2010. 4327

[10] R. Fan, K. Chang, C. Heish, X. Wang, and C. Lin. LIBLINEAR: A library for large linear classification. *JMLR*, 9:1871–1874, 2008. 4326

[11] C. Galleguillos, B. Babenko, A. Rabinovich, and S. Belongie. Weakly supervised object localization with stable segmentations. *Proc. ECCV*, pages 193–207, 2008. 4322

[12] M. Grundmann, V. Kwatra, M. Han, and I. Essa. Efficient hierarchical graph-based video segmentation. *Proc. CVPR*, pages 2141–2148, 2010. 4322

[13] J. Hammersley and P. Clifford. Markov fields on graphs and lattices. *Unpublished manuscript*, 1971. Available at http://www.statslab.cam.ac.uk/~grg/books/hammfest/hamm-cliff.pdf. 4325

[14] G. Hartmann, M. Grundmann, J. Hoffman, D. Tsai, V. Kwatra, O. Madani, S. Vijayanarasimhan, I. Essa, J. Rehg, and R. Sukthankar. Weakly supervised learning of object segmentations from web-scale video. In *Proc. ECCV*, pages 198–208, 2012. 4321, 4322, 4324, 4326, 4327

[15] Y. Huang, Q. Liu, and D. N. Metaxas. Video object segmentation by hypergraph cut. *Proc. CVPR*, pages 1738–1745, 2009. 4322

[16] A. Jain, S. Chatterjee, and R. Vidal. Coarse-to-fine semantic video segmentation using supervoxel trees. *Proc. ICCV*, 2013. 4325

[17] W. Liu, J. Wang, R. Ji, Y.-G. Jiang, and S.-F. Chang. Supervised hashing with kernels. *Proc. CVPR*, pages 2074–2081, 2012. 4323

[18] X. Liu, M. Song, D. Tao, Z. Liu, C. Chen, and J. Bu. Semi-supervised node splitting for random forest construction. *Proc. CVPR*, pages 492–499, 2013. 4325

[19] T. Ojala, M. Pietikainen, and T. Maepaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24:971–987, 2002. 4325

[20] R. Potts. Some generalized order-disorder transformations. *Proc. Mathematical*, pages 106–109, 1952. 4325

[21] A. Prest, C. Leistner, J. Civera, C. Schmid, and V. Ferrari. Learning object class detectors from weakly annotated video. *Proc. CVPR*, pages 3282–3289, 2012. 4327

[22] S. Raza, M. Grundmann, and I. Essa. Geometric context from videos. *Proc. CVPR*, pages 3081–3088, 2013. 4322

[23] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22:888–905, 2000. 4322

[24] J. Shotton, M. Johnson, and R. Cipolla. Segmantic texton forests for image categorization and segmentation. *Proc. CVPR*, pages 1–8, 2008. 4322

[25] P. Siva, C. Russell, and T. Xiang. In defence of negative mining for annotating weakly labelled data. *Proc. ECCV*, pages 594–608, 2012. 4326, 4327

[26] M. Szummer, P. Kohli, and D. Hoiem. Learning CRFs using graph cuts. *Proc. ECCV*, pages 582–595, 2008. 4325

[27] K. Tang, R. Sukthankar, J. Yagnik, and L. Fei-Fei. Discriminative segment annotation in weakly labeled video. *Proc. CVPR*, pages 2483–2490, 2013. 4321, 4322, 4324, 4326, 4327

[28] M. Vasconcelos, G. Carneiro, and N. Vasconcelos. Weakly supervised top-down image segmentation. *Proc. CVPR*, pages 1001–1006, 2006. 4322

[29] V.badrinarayanan, F. Galasso, M. Johnson, and R. Cipolla. Label and propagation in video sequences. *Proc. CVPR*, pages 3265–3272, 2010. 4322

[30] A. Vezhnevets and J. Buhmann. Towards weakly supervised semantic segmentation by means of multiple instance and multitask learning. *Proc. CVPR*, pages 3249–3256, 2010. 4321, 4322, 4326

[31] A. Vezhnevets, V. Ferrari, and J. Buhmann. Weakly supervised semantic segmentation with a multi-image model. *Proc. ICCV*, pages 643–650, 2011. 4321, 4322, 4326, 4327

[32] A. Vezhnevets, V. Ferrari, and J. Buhmann. Weakly supervised structured output learning for semantic segmentation. *Proc. CVPR*, pages 845–852, 2012. 4321, 4322

[33] J. Wang, S. Kumar, and S. Chang. Semi-supervised hashing for scalable image retrieval. *Proc. CVPR*, pages 3424–3431, 2010. 4323

[34] C. Xu and J. Carso. Evaluation of super-voxel methods for early video processing. *Proc. CVPR*, pages 1202–1209, 2012. 4325

[35] C. Xu, C. Xiong, and J. Corso. Streaming hierarchical video segmentation. *Proc. ECCV*, pages 626–639, 2012. 4322, 4325

[36] L. Zhang, Y. Gao, R. Ji, L. Ke, and J. Shen. Representative discovery of structure cues for weakly-supervised image segmentation. *IEEE Trans. Multimedia*, 2013. 4321, 4322

[37] L. Zhang, M. Song, Z. Liu, X. Liu, J. Bu, and C. Chen. Probabilistic graphlet cut: Exploiting spatial structure cue for weakly supervised image segmentation. *Proc. CVPR*, pages 1908–1915, 2013. 4322