Discriminative Segment Annotation in Weakly Labeled Video

Kevin Tang[1,2] Rahul Sukthankar[1] Jay Yagnik[1] Li Fei-Fei[2]
kdtang@cs.stanford.edu rahuls@cs.cmu.edu jyagnik@google.com feifeili@cs.stanford.edu
[1]Google Research [2]Computer Science Department, Stanford University
https://sites.google.com/site/segmentannotation/

Abstract

The ubiquitous availability of Internet video offers the vision community the exciting opportunity to directly learn localized visual concepts from real-world imagery. Unfortunately, most such attempts are doomed because traditional approaches are ill-suited, both in terms of their computational characteristics and their inability to robustly contend with the label noise that plagues uncurated Internet content. We present CRANE, a weakly supervised algorithm that is specifically designed to learn under such conditions. First, we exploit the asymmetric availability of real-world training data, where small numbers of positive videos tagged with the concept are supplemented with large quantities of unreliable negative data. Second, we ensure that CRANE is robust to label noise, both in terms of tagged videos that fail to contain the concept as well as occasional negative videos that do. Finally, CRANE is highly parallelizable, making it practical to deploy at large scale without sacrificing the quality of the learned solution. Although CRANE is general, this paper focuses on segment annotation, where we show state-of-the-art pixel-level segmentation results on two datasets, one of which includes a training set of spatiotemporal segments from more than 20,000 videos.

1. Introduction

The ease of authoring and uploading video to the Internet creates a vast resource for computer vision research, particularly because Internet videos are frequently associated with semantic tags that identify visual concepts appearing in the video. However, since tags are not spatially or temporally localized within the video, such videos cannot be directly exploited for training traditional supervised recognition systems. This has stimulated significant recent interest in methods that learn localized concepts under weak supervision [11, 16, 20, 25]. In this paper, we examine the problem of generating pixel-level concept annotations for weakly labeled video.

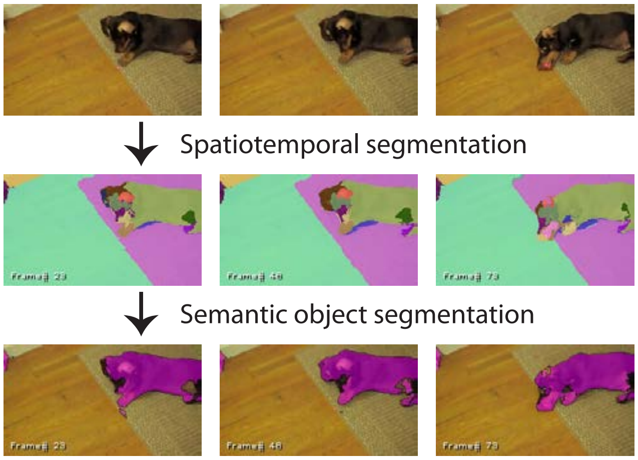

Figure 1. Output of our system. Given a weakly tagged video (e.g., "dog") [top], we first perform unsupervised spatiotemporal segmentation [middle]. Our method identifies segments that correspond to the label to generate a semantic segmentation [bottom].

To make our problem more concrete, we provide a rough pipeline of the overall process (see Fig. 1). Given a video weakly tagged with a concept, such as "dog", we process it using a standard unsupervised spatiotemporal segmentation method that aims to preserve object boundaries [3, 10, 15]. From the video-level tag, we know that some of the segments correspond to the "dog" concept while most probably do not. Our goal is to classify each segment within the video either as coming from the concept "dog", which we denote as *concept segments*, or not, which we denote as *background segments*. Given the varied nature of Internet videos, we cannot rely on assumptions about the relative frequencies or spatiotemporal distributions of segments from the two classes, neither within a frame nor across the video; nor can we assume that each video contains a single instance of the concept. For instance, neither the dog in Fig. 1 nor most of the objects in Fig. 10 would be separable from the complex background by unsupervised methods.

There are two settings for addressing the segment annotation problem, which we illustrate in Fig. 2. The first scenario, which we term *transductive segment annotation* (TSA), is studied in [23]. This scenario is closely related to automatically annotating a weakly labeled dataset. Here, the test videos that we seek to annotate are compared against a large amount of negative segments (from videos not tagged with the concept) to enable a direct discriminative separation of the test video segments into two classes. The second scenario, which we term *inductive segment annotation* (ISA), is studied in [11]. In this setting, a segment classifier is trained using a large quantity of weakly labeled segments from both positively- and negatively-tagged videos. Once trained, the resulting classifier can be applied to any test video (typically not in the original set). We observe that the TSA and ISA settings parallel the distinction between transductive and inductive learning, since the test instances are available during training in the former but not in the latter. Our proposed algorithm, Concept Ranking According to Negative Exemplars (CRANE), can operate under either scenario and we show experimental results demonstrating its clear superiority over previous work under both settings.

Our contributions can be organized into three parts.

1. We present a unified interpretation under which a broad class of weakly supervised learning algorithms can be analyzed.
2. We introduce CRANE, a straightforward and effective discriminative algorithm that is robust to label noise and highly parallelizable. These properties of CRANE are extremely important, as such algorithms must handle large amounts of video data and spatiotemporal segments.
3. We introduce spatiotemporal segment-level annotations for a subset of the YouTube-Objects dataset [20], and present a detailed analysis of our method compared to other methods on this dataset for the transductive segment annotation scenario. To promote research into this problem, we make our annotations freely available.[1] We also compare CRANE directly against [11] on the inductive segment annotation scenario and demonstrate state-of-the-art results.

2. Related Work

Several methods have recently been proposed for high-quality, unsupervised spatiotemporal segmentation of videos [3, 10, 15, 30, 31]. The computational efficiency of some of these approaches [10, 31] makes it feasible to segment large numbers of Internet videos. Several recent works have leveraged spatiotemporal segments for a variety of tasks in video understanding, including event detection [12], human motion volume generation [17], human activity recognition [2], and object segmentation [11, 13]. Drawing inspiration from these, we also employ such segments as a core representation in our work.

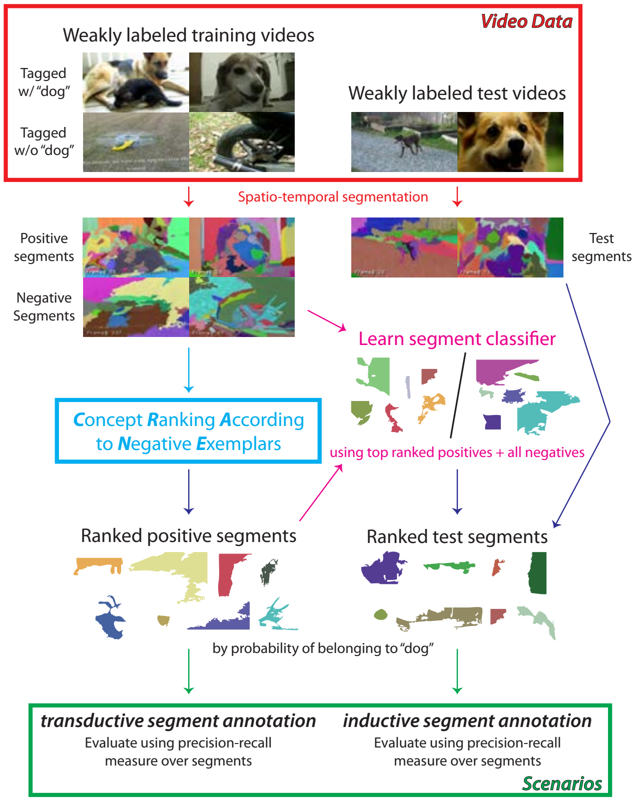
Video Data
Weakly labeled training videos
Tagged w/ "dog"
Tagged w/o "dog"
Weakly labeled test videos
Spatio-temporal segmentation
Positive segments
Negative Segments
Test segments
Learn segment classifier
Concept Ranking According to Negative Exemplars
using top ranked positives + all negatives
Ranked positive segments
Ranked test segments
by probability of belonging to "dog"
transductive segment annotation
Evaluate using precision-recall measure over segments
inductive segment annotation
Evaluate using precision-recall measure over segments
Scenarios

Figure 2. Overview of transductive and inductive segment annotation. In the former (TSA), the proposed algorithm (CRANE) is evaluated on weakly labeled training data; in the latter (ISA), we train a classifier and evaluate on a disjoint test set. TSA and ISA have parallels to transductive and inductive learning, respectively.

Lee *et al.* [13] perform object segmentation on unannotated video sequences. Our approach is closer to that of Hartmann *et al.* [11], where object segmentations are generated on weakly labeled video data. Whereas [11] largely employ variants on standard supervised methods (e.g., linear classifiers and multiple-instance learning), we propose a new way of thinking about this weakly supervised problem that leads to significantly superior results.

Discriminative segment annotation from weakly labeled data shares similarities with Multiple Instance Learning (MIL), on which there has been considerable research (e.g., [5, 28, 32, 33]). In MIL, we are given labeled bags of instances, where a positive bag contains at least one positive instance, and a negative bag contains no positive instances. MIL is more constrained than our scenario, since these guarantees may not hold due to label noise (which is typically present in video-level tags). In particular, algorithms must contend with positive videos that actually con-

[1] Annotations and additional details are available at the project website: https://sites.google.com/site/segmentannotation/.

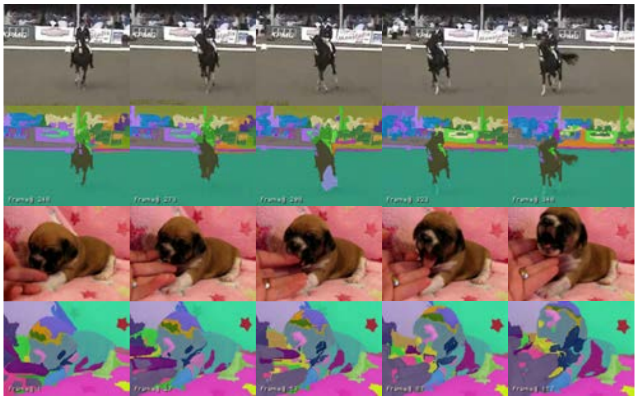

Figure 3. Spatiotemporal segments computed on "horse" and "dog" video sequences using [10]. Segments with the same color correspond across frames in the same sequence.

tain no concept segments as well as rare cases where some concept segments appear in negative videos.

There is increasing interest in exploring the idea of learning visual concepts from a combination of weakly supervised images and weakly supervised video [1, 6, 14, 19, 21, 26]. Most applicable to our problem is recent work that achieves state-of-the-art results on bounding box annotation in weakly labeled 2D images [23]. We show that this "negative mining" method can also be applied to segment annotation. Direct comparisons show that CRANE outperforms negative mining and is more robust to label noise.

3. Weakly Supervised Segment Annotation

As discussed earlier, we start with spatiotemporal segments for each video, such as those shown in Fig. 3. Each segment is a spatiotemporal (3D) volume that we represent as a point in a high-dimensional feature space using a set of standard features computed over the segment.

More formally, for a particular concept c, we are given a dataset $\{\langle s_1, y_1 \rangle, ..., \langle s_N, y_N \rangle\}$, where s_i is segment i, and $y_i \in \{-1, 1\}$ is the label for segment i, with the label being positive if the segment was extracted from a video with concept c as a weak label, and negative otherwise. We denote the set $\mathcal{P}$ to be the set of all instances with a positive label, and similarly $\mathcal{N}$ to be the set of all negative instances. Since our negative data was weakly labeled with concepts other than c, we can assume that the segments labeled as negative are (with rare exceptions) correctly labeled. Our task then is to determine which of the positive segments $\mathcal{P}$ are concept segments, and which are background segments.

We present a generalized interpretation of transductive segment annotation, which leads to a family of methods that includes several common methods and previous works [23]. Consider the pairwise distance matrix (in the high-dimensional feature space) between all of the segments s_i from both the positive and negative videos, for a particular concept c. Across the rows and columns, we order the segments from $\mathcal{P}$ first, followed by those from $\mathcal{N}$. Within $\mathcal{P}$, we further order the concept segments $\mathcal{P}_c \subset \mathcal{P}$ first, followed by the background segments $\mathcal{P}_b = \mathcal{P} \setminus \mathcal{P}_c$. This distance matrix is illustrated in Fig. 4. The blocks A, B and C correspond to intra-class distances among segments from $\mathcal{P}_c$, $\mathcal{P}_b$, and $\mathcal{N}$, respectively. The block circumscribing A and B corresponds to the distances among $\mathcal{P}$. Note that A and B are hidden from the algorithm, since determining the membership of $\mathcal{P}_c$ is the goal of TSA. We can now analyze a variety of weakly supervised approaches in this framework.

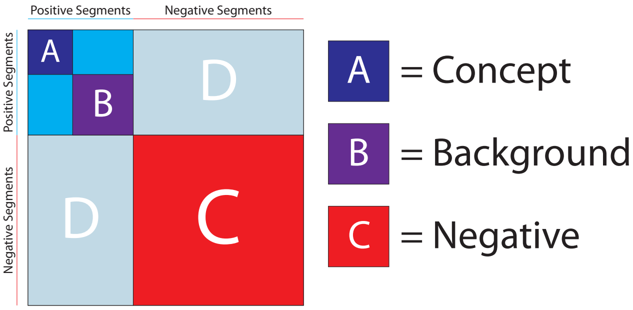

Figure 4. Visualization of pairwise distance matrix between segments for weakly supervised annotation. See text for details.

Rather than solely studying TSA as the problem of partitioning $\mathcal{P}$, we find it fruitful to also consider the related problem of *ranking* the elements of $\mathcal{P}$ in decreasing order of a score, $S(s_i)$ such that top-ranked elements correspond to $\mathcal{P}_c$; thresholding at a particular rank generates a partition.

Co-segmentation/Clustering. Co-segmentation [27] exploits the observation that concept segments across videos are similar, but that background segments are diverse. The purest variants of this approach are unsupervised and do not require $\mathcal{N}$ and can operate solely on the top-left 2×2 submatrix. The hope is that the concept segments form a dominant cluster/clique in feature space.

Kernel density estimation for $\mathcal{N}$. This principled approach to weakly supervised learning exploits the insight that the (unknown) distribution of background segments $\mathcal{P}_b$ must be similar to the (known) distribution of negative segments $\mathcal{N}$, since the latter consists almost entirely of background segments. Accordingly, we construct a non-parametric model of the probability density $P_{\mathcal{N}}(x)$ generated from the latter (block C) and employ it as a proxy for the former (block B). Then, elements from $\mathcal{P}$ that lie in high-density regions of $P_{\mathcal{N}}(.)$ can be assumed to come from $\mathcal{P}_b$, while those in low-density regions are probably the concepts $\mathcal{P}_c$ that we seek. A natural algorithm for TSA is thus to rank the elements $s_i \in \mathcal{P}$ according to $P_{\mathcal{N}}(s_i)$.

In practice, we estimate $P_{\mathcal{N}}$ using kernel density estimation, with a Gaussian kernel whose σ is determined using cross-validation so as to maximize the log likelihood of generating $\mathcal{N}$. In our interpretation, this corresponds to building a generative model according to the information in block C of the distance matrix, and scoring segments according to:

$$S_{\text{KDE}}(s_i) = -P_{\mathcal{N}}(s_i) = -\frac{1}{|\mathcal{N}|}\sum_{z\in\mathcal{N}} N\Big(\text{dist}(s_i, z); \sigma^2\Big), \tag{1}$$

where $N(\cdot;\sigma^2)$ denotes a zero-mean multivariate Gaussian with isotropic variance of σ^2.

Supervised discriminative learning with label noise. Standard fully supervised methods, such as Support Vector Machines (SVM), learn a discriminative classifier to separate positive from negative data, given instance-level labels. Such methods can be shoehorned into the weakly supervised setting of segment annotation by propagating video-level labels to segments. In other words, we learn a discriminative classifier to separate $\mathcal{P}$ from $\mathcal{N}$, or the upper 2×2 submatrix vs. block C. Unfortunately, since $\mathcal{P} = \mathcal{P}_c \cup \mathcal{P}_b$, this approach treats the background segments from positively tagged videos, $\mathcal{P}_b$ (which are typically the majority), as label noise. Nonetheless, such approaches have been reported to perform surprisingly well [11], where linear SVMs trained with label noise achieve competitive results. This may be because the limited capacity of the classifier is unable to separate $\mathcal{P}_b$ from $\mathcal{N}$ and therefore focuses on separating $\mathcal{P}_c$ from $\mathcal{N}$. In our experiments, methods that tackle weakly labeled segment annotation from a more principled perspective significantly outperform these techniques.

Negative Mining (MIN). Siva *et al.*'s negative mining method [23], which we denote as MIN, can be interpreted as a discriminative method that operates on block D of the matrix to identify $\mathcal{P}_c$. Intuitively, distinctive concept segments are identified as those among $\mathcal{P}$ whose nearest neighbor among $\mathcal{N}$ is as far as possible. Operationally, this leads to the following score for segments:

$$S_{\text{MIN}}(s_i) = \min_{t\in\mathcal{N}} \Big(\text{dist}(s_i, t)\Big). \tag{2}$$

Following this perspective on how various weakly supervised approaches for segment annotations relate through the distance matrix, we detail our proposed algorithm, CRANE.

4. Proposed Method: CRANE

Like MIN, our method, CRANE, operates on block D of the matrix, corresponding to the distances between weakly tagged positive and negative segments. Unlike MIN, CRANE iterates through the segments in $\mathcal{N}$, and each such negative instance penalizes nearby segments in $\mathcal{P}$. The intuition is that concept segments in $\mathcal{P}$ are those that are far from negatives (and therefore less penalized). While one can envision several algorithms that exploit this theme, the simplest variant of CRANE can be characterized by the following segment scoring function:

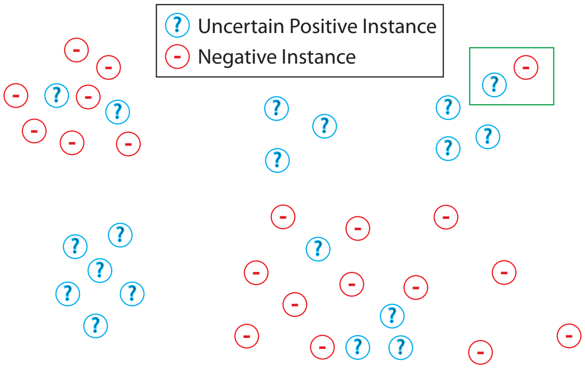

Figure 5. Intuition behind CRANE. Positive instances are less likely to be concept segments if they are near many negatives. The green box contrasts CRANE with MIN [23] as discussed in text.

$$S_{\text{CRANE}}(s_i) = -\sum_{z\in\mathcal{N}} \mathbf{1}\Big[s_i = \arg\min_{t\in\mathcal{P}} \Big(\text{dist}(t, z)\Big)\Big] \cdot f_{\text{cut}}\Big(\text{dist}(s_i, z)\Big), \tag{3}$$

where $\mathbf{1}(\cdot)$ denotes the indicator function and $f_{\text{cut}}(\cdot)$ is a cutoff function over an input distance.

Fig. 5 illustrates the intuition behind CRANE. Background segments in positive videos tend to fall near one or more segments from negative videos (in feature space). The nearest neighbor to every negative instance is assigned a penalty $f_{\text{cut}}(.)$. Consequently, such segments are ranked lower than other positives. Since concept segments are rarely the closest to negative instances, they are typically ranked higher. Fig. 5 also shows how CRANE is more robust than MIN [23] to label noise among negative videos. Consider the points in the green box shown at the top right of the figure. Here, the unknown segment, s_i, is very close to a negative instance that may have come from an incorrectly tagged video. This single noisy instance will cause MIN to irrecoverably reject s_i. By contrast, CRANE will just assign s_i a small penalty for its proximity and in the absence of corroborating evidence from other negative instances, s_i's rank will not change significantly.

Before detailing the specifics of how we apply CRANE to transductive and inductive segment annotation tasks, we discuss some properties of the algorithm that make it particularly suitable to practical implementations. First, as mentioned above, CRANE is robust to noise, whether from incorrect labels or distorted features, confirmed in controlled experiments (see Section 5.1). Second, CRANE is explicitly designed to be parallelizable, enabling it to employ large numbers of negative instances. Motivated by Siva *et* *al.* [23]'s observation regarding the abundance of negative data, our proposed approach enforces independence among negative instances (i.e., explicitly avoids using the data from block C of the distance matrix). This property enables CRANE's computation to be decomposed over a large number of machines simply by replicating the positive instances, partitioning the (much larger) negative instances, and trivially aggregating the resulting scores.

4.1. Application to transductive segment annotation

Applying CRANE to transductive segment annotation is straightforward. We generate weakly labeled positive and negative instances for each concept. Then we use CRANE to rank all of the segments in the positive set according to this score. Thresholding the list at a particular rank creates a partitioning into P_c and P_b; sweeping the threshold generates the precision/recall curves shown in Fig. 6.

4.2. Application to inductive segment annotation

In the inductive segment annotation task, for each concept, we are given a large number of weakly tagged positive and negative videos, from which we learn a set of segment-level classifiers that can be applied to arbitrary weakly tagged test videos. Inductive segment annotation can be decomposed into a two-stage problem. The first stage is identical to TSA. In the second stage, the most confident predictions for concept segments (from the first stage) are treated as segment-level labels. Using these and our large set of negative instances, we train a standard fully supervised classifier. To evaluate the performance of ISA, we apply the trained classifier to a disjoint test set and generate precision/recall curves, such as those shown in Fig. 8.

5. Experiments

To evaluate the different methods, we score each segment in our test videos, rank segments in decreasing order of score and compute precision/recall curves. As discussed above, the test videos for TSA are available during training, whereas those for ISA are disjoint from the training videos.

5.1. Transductive segment annotation (TSA)

To evaluate transductive segment annotation, we use the YouTube-Objects (YTO) dataset [20], which consists of videos collected for 10 of the classes from the PASCAL Visual Objects Challenge [8]. We generate a groundtruthed test set by manually annotating the first shot from each video with segment-level object annotations, resulting in a total of 151 shots with a total of 25,673 frames (see Table 1) and 87,791 segments. We skip videos for which the object did not occur in the first shot and shots with severe undersegmentation problems. Since there is increasing interest in training image classifiers using video data [20, 24], our hope is to identify methods that can "clean" weakly supervised video to generate suitable data for training supervised classifiers for image challenges such as PASCAL VOC.

Class	Shots	Frames	Class	Shots	Frames
Aeroplane	9	1423	Cow	20	2978
Bird	6	1206	Dog	27	3803
Boat	17	2779	Horse	17	3990
Car	8	601	Motorbike	11	829
Cat	18	4794	Train	18	3270
Total Shots		151	Total Frames		25673

Table 1. Details for our annotations on the YouTube-Objects dataset [20]. Note that each shot comes from a different video, as we do not annotate multiple shots in the same video.

Implementation details. We represent each segment using the following set of features: RGB color histograms quantized over 20 bins, histograms of local binary patterns computed on 5×5 patches [18, 29], histograms of dense optical flow [4], heat maps computed over an 8×6 grid to represent the (x, y) shape of each segment (averaged over time), and histograms of quantized SIFT-like local descriptors extracted densely within each segment. For negative data, we sample 5000 segments from videos tagged with other classes; our experiments show that additional negative data increases computation time but does not significantly affect results for any of the methods on this dataset.

We use the L2 distance for the distance function in relevant methods, and for the cutoff function in CRANE, we simply use a constant, $f_{\text{cut}}(\cdot) = 1$. Experiments with cutoff functions such as step, ramp and Gaussian show that the constant performs just as well and requires no parameters.

Direct comparisons. We compare CRANE against several methods. MIL refers to Multiple Instance Learning, the standard approach for problems similar to our scenario. In our experiments, we use the MILBoost algorithm with ISR criterion [28], and sparse boosting with decision stumps [7] as the base classifier. MIN refers to the method of [23], which uses the minimum distance for each positive instance as the score for the instance. KDE refers to Kernel Density Estimation, which estimates the probability distribution of the negatives, and then computes the probability that each positive instance was generated from this distribution.

Discussion. Fig. 6 shows that our method outperforms all other methods in overall precision/recall. In particular, we perform much better for the "aeroplane", "dog", "horse", and "train" classes. Interestingly, for the "cat" class, MIL performs very well whereas all other methods do poorly. By visualizing the segments (see Fig. 7), we see that in many videos, the cat and background segments are very similar in appearance. MIL is able to focus on these minor differences while the others do not. MIN [23] performs second best on this task after CRANE. However, because it only considers

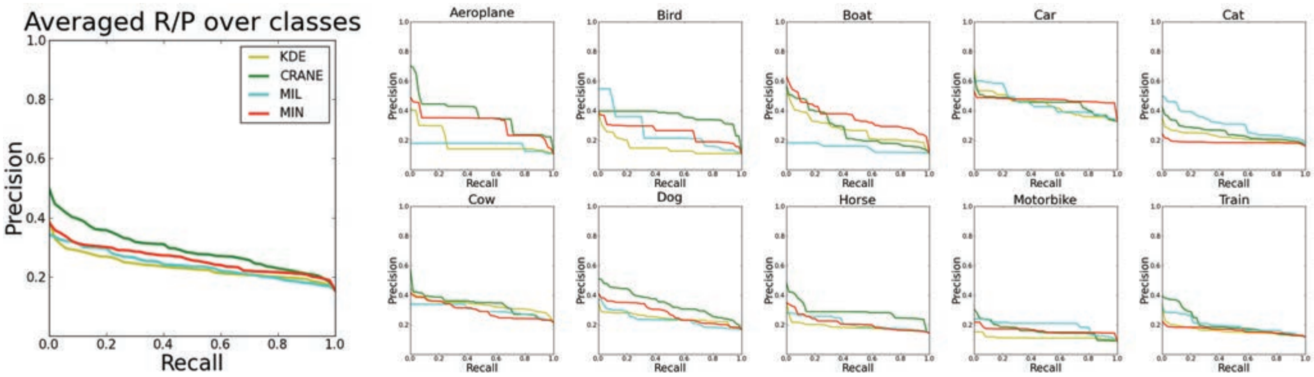

Figure 6. Direct comparison of several approaches for transductive segment annotation on the YouTube-Objects dataset [20].

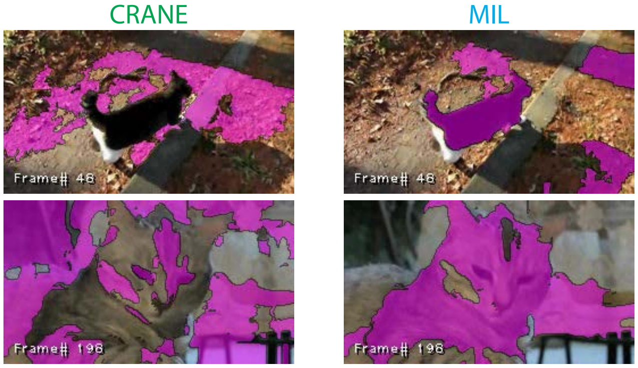

Figure 7. Visualizations of instances for the "cat" class where MIL is better able to distinguish between the similar looking concept and background segments (see text for details).

the minimum distance from a positive instance to a negative instance, it is more susceptible to label noise.

The transductive segment annotation scenario is useful for directly comparing various weakly supervised learning methods in a classifier-independent manner. However, TSA is of limited practical use as it requires that each segment from every input video be compared against the negative data. By contrast, ISA assumes that once a segment-level concept model has been learned (using sufficient data to span the concept's intra-class variability), the model can be applied relatively efficiently to arbitrary input videos.

5.2. Inductive segment annotation (ISA)

For the task of inductive segment annotation, where we learn a segment-level classifier from weakly labeled video, we use the dataset introduced by [11], as this dataset contains a large number of weakly labeled videos and deals exactly with this task. This dataset consists of 20,000 Internet videos from 8 classes: "bike", "boat", "card", "dog", "helicopter", "horse", "robot", and "transformer". Additional videos from several other tags are used to increase the set of negative background videos. These videos are used for training, and a separate, disjoint set of test videos from these 8 concept classes is used for evaluation.

Implementation details. Due to the computational limitations of the MIL baseline, we limit the training set to 200,000 segments, equally divided among samples from $\mathcal{P}$ and $\mathcal{N}$. For segment features, we use RGB color histograms and histograms of local binary patterns. For both CRANE and MIN, we retain the top 20% of the ranked segments from $\mathcal{P}$ as positive training data for the second stage segment classifier. To simplify direct comparisons, we use k-nearest neighbor (kNN) as the second-stage classifier, with k=20 and probabilistic output for x generated as the ratio to closest negative vs. closest positive: $\min_{n\in\mathcal{N}} ||x-n|| / \min_{p\in\mathcal{P}} ||x-p||$.

Direct comparisons. In addition to several of the stronger methods from the TSA task, we add two baselines for the ISA task: (1) kNN denotes the same second-stage classifier, but using all of the data $\mathcal{P} \cup \mathcal{N}$; (2) SVM refers to a linear support vector machine implemented using LIBLINEAR [9] that was reported to do well by [11] on their task.

Discussion. Fig. 8 shows that CRANE significantly outperforms the others in overall precision/recall and dominates in most of the per-class comparisons. In particular, we see strong gains (except on "dog") vs. MIL, which is important because [11] was unable to show significant gains over MIL on this dataset. SVM trained with label noise performs worst, except for a few low-recall regions where SVM does slightly better, but no method performs particularly well.

Fig. 9 (top) examines how CRANE's average precision on ISA varies with the fraction of retained segments. As expected, if we retain too few segments, we do not span the intra-class variability of the target concept; conversely, retaining too many concepts risks including background segments and consequently corrupting the learned classifier. Fig. 9 (bottom) shows the effect of additional training data (with 20% retained segments). We see that average precision improves quickly with training data and plateaus

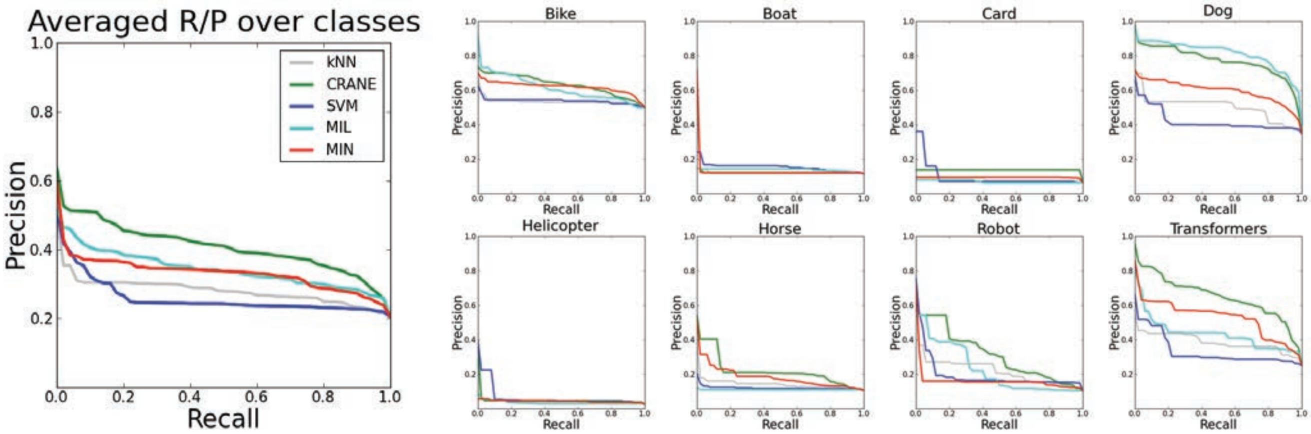

Figure 8. Direct comparison of several methods for inductive segment annotation using the object segmentation dataset [11].

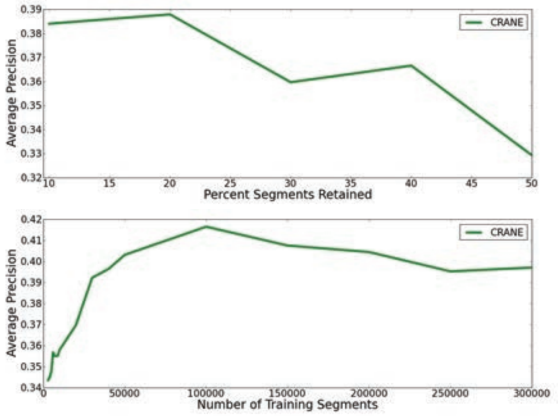

Figure 9. Average precision as we vary CRANE's fraction of retained segments [top] and number of training segments [bottom].

around 0.4 once we exceed 100,000 training segments.

Fig. 10 shows example successes and failures for CRANE under both TSA and ISA settings. We stress that these results (unlike those in [11]) are the raw outputs of independent segment-level classification and employ no intra-segment post-processing to smooth labels. Observations on successes: we segment multiple non-centered objects (top-left), which is difficult for GrabCut-based methods [22]; we highlight the horse but not the visually salient ball, improving over [11]; we find the speedboat but not the moving water. CRANE can occasionally fail in clutter (top right) or when segmentations are of low quality (cruise ship + water).

6. Conclusion

We introduce CRANE, a surprisingly simple yet effective algorithm for annotating spatiotemporal segments from video-level labels. We also present a generalized interpretation based on the distance matrix that serves as a taxonomy for weakly supervised methods and provides a deeper understanding of this problem. We describe two related scenarios of the segment annotation problem (TSA and ISA) and present comprehensive experiments on published datasets. CRANE outperforms the recent methods [11, 23] as well as our baselines on both TSA and ISA tasks.

There are many possible directions for future work. In particular, CRANE is only one of a family of methods that exploit distances between weakly labeled instances for discriminative ranking and classification. Much of the distance matrix remains to be fully leveraged and understanding how best to use the other blocks is an interesting direction.

Acknowledgments. This research was conducted during K. Tang's internship at Google. We thank T. Dean, M. Grundmann, J. Hoffman, A. Kovashka, V. Kwatra, K. Murphy, O. Madani, D. Ross, M. Ruzon, M. Segal, J. Shlens, G. Toderici, D. Tsai, and S. Vijayanarasimhan for valuable code and discussions. We also thank J. Deng, D. Held, and V. Ramanathan for helpful comments on the paper.

References

[1] K. Ali, D. Hasler, and F. Fleuret. FlowBoost—Appearance learning from sparsely annotated video. In *CVPR*, 2011. 3

[2] W. Brendel and S. Todorovic. Learning spatiotemporal graphs of human activities. In *ICCV*, 2011. 2

[3] T. Brox and J. Malik. Object segmentation by long term analysis of point trajectories. In *ECCV*, 2010. 1, 2

[4] R. Chaudhry et al. Histograms of oriented optical flow and Binet-Cauchy kernels on nonlinear dynamical systems for the recognition of human actions. In *CVPR*, 2009. 5

[5] Y. Chen, J. Bi, and J. Wang. MILES: Multiple-instance learning via embedded instance selection. *PAMI*, 28(12), 2006. 2

[6] T. Deselaers, B. Alexe, and V. Ferrari. Localizing objects while learning their appearance. In *ECCV*, 2010. 3

[7] J. C. Duchi and Y. Singer. Boosting with structural sparsity. In *ICML*, 2009. 5

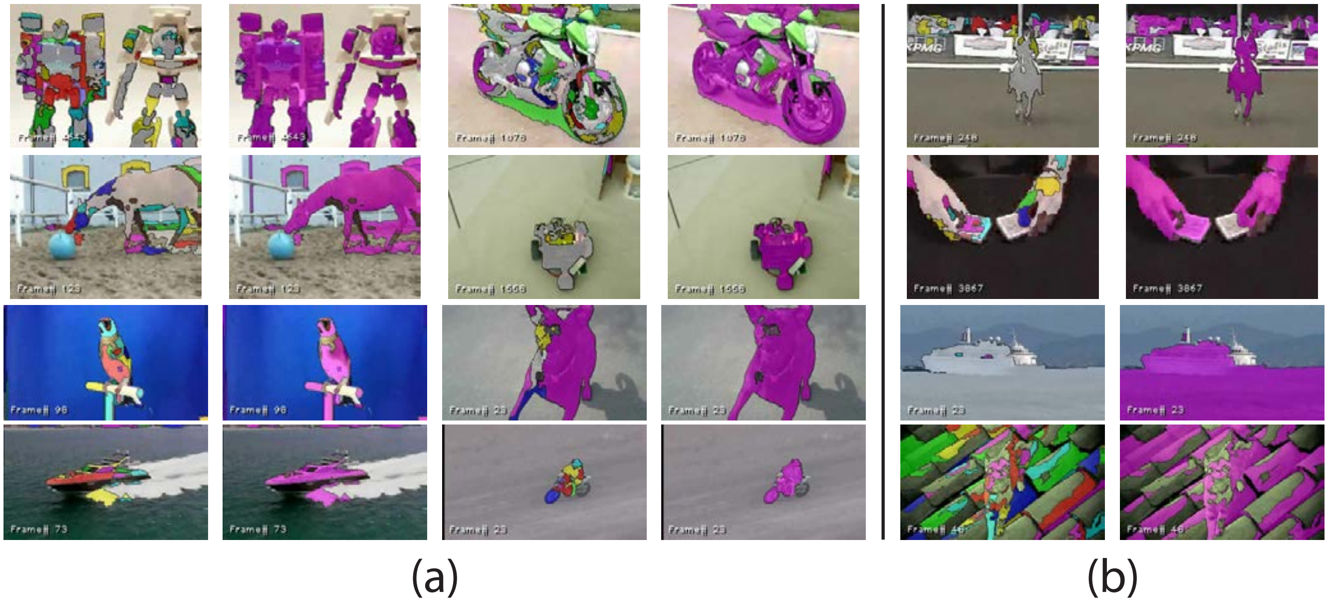

Figure 10. Object segmentations obtained using CRANE. The top two rows are obtained for the ISA task on the dataset introduced by [11]. The bottom two rows are obtained for the TSA task on the YouTube-Objects dataset [20]. In each pair, the left image shows the original spatiotemporal segments and the right shows the output. (a) Successes; (b) Failures.

[8] M. Everingham et al. The Pascal visual object classes (VOC) challenge. *IJCV*, 2010. 5

[9] R.-E. Fan et al. LIBLINEAR: A library for large linear classification. *JMLR*, 9, 2008. 6

[10] M. Grundmann, V. Kwatra, M. Han, and I. A. Essa. Efficient hierarchical graph-based video segmentation. In *CVPR*, 2010. 1, 2, 3

[11] G. Hartmann, M. Grundmann, J. Hoffman, D. Tsai, V. Kwatra, O. Madani, S. Vijayanarasimhan, I. A. Essa, J. M. Rehg, and R. Sukthankar. Weakly supervised learning of object segmentations from web-scale video. In *ECCV Workshop on Vision in Web-Scale Media*, 2012. 1, 2, 4, 6, 7, 8

[12] Y. Ke, R. Sukthankar, and M. Hebert. Event detection in crowded videos. In *ICCV*, 2007. 2

[13] Y. Lee, J. Kim, and K. Grauman. Key-segments for video object segmentation. In *ICCV*, 2011. 2

[14] C. Leistner et al. Improving classifiers with unlabeled weakly-related videos. In *CVPR*, 2011. 3

[15] J. Lezama, K. Alahari, J. Sivic, and I. Laptev. Track to the future: Spatio-temporal video segmentation with long-range motion cues. In *CVPR*, 2011. 1, 2

[16] J. C. Niebles, C.-W. Chen, and L. Fei-Fei. Modeling temporal structure of decomposable motion segments for activity classification. In *ECCV*, 2010. 1

[17] J.-C. Niebles, B. Han, A. Ferencz, and L. Fei-Fei. Extracting moving people from internet videos. In *ECCV*, 2008. 2

[18] T. Ojala, M. Pietikainen, and D. Harwood. Performance evaluation of texture measures with classification based on Kullback discrimination of distributions. In *ICPR*, 1994. 5

[19] B. Ommer, T. Mader, and J. Buhmann. Seeing the objects behind the dots: Recognition in videos from a moving camera. *IJCV*, 83(1), 2009. 3

[20] A. Prest, C. Leistner, J. Civera, C. Schmid, and V. Ferrari. Learning object class detectors from weakly annotated video. In *CVPR*, 2012. 1, 2, 5, 6, 8

[21] D. Ramanan, D. Forsyth, and K. Barnard. Building models of animals from video. *PAMI*, 28(8), 2006. 3

[22] C. Rother et al. GrabCut: interactive foreground extraction using iterated graph cuts. *Trans. Graphics*, 23(3), 2004. 7

[23] P. Siva, C. Russell, and T. Xiang. In defence of negative mining for annotating weakly labelled data. In *ECCV*, 2012. 2, 3, 4, 5, 7

[24] K. Tang et al. Shifting weights: Adapting object detectors from image to video. In *NIPS*, 2012. 5

[25] K. Tang, L. Fei-Fei, and D. Koller. Learning latent temporal structure for complex event detection. In *CVPR*, 2012. 1

[26] A. Vezhnevets, V. Ferrari, and J. M. Buhmann. Weakly supervised structured output learning for semantic segmentation. In *CVPR*, 2012. 3

[27] S. Vicente, V. Kolmogorov, and C. Rother. Cosegmentation revisited: Models and optimization. In *ECCV*, 2010. 3

[28] P. Viola, J. Platt, and C. Zhang. Multiple instance boosting for object detection. In *NIPS*, 2005. 2, 5

[29] X. Wang, T. X. Han, and S. Yan. An HOG-LBP human detector with partial occlusion handling. In *ICCV*, 2009. 5

[30] J. Xiao and M. Shah. Motion layer extraction in the presence of occlusion using graph cuts. *PAMI*, 27(10), 2005. 2

[31] C. Xu, C. Xiong, and J. Corso. Streaming hierarchical video segmentation. In *ECCV*, 2012. 2

[32] Z.-J. Zha et al. Joint multi-label multi-instance learning for image classification. In *CVPR*, 2008. 2

[33] Z.-H. Zhou and M.-L. Zhang. Multi-instance multi-label learning with application to scene classification. In *NIPS*, 2007. 2