Active Frame, Location, and Detector Selection for Automated and Manual Video Annotation

Vasiliy Karasev
UCLA Vision Lab
University of California
Los Angeles, CA 90095
vasiliy.karasev@ucla.edu

Avinash Ravichandran
Amazon
500 Boren Ave N,
Seattle, WA 98109
avinash@cs.ucla.edu

Stefano Soatto
UCLA Vision Lab
University of California
Los Angeles, CA 90095
soatto@cs.ucla.edu

http://vision.ucla.edu/activeselection/

Abstract

We describe an information-driven active selection approach to determine which detectors *to deploy at* which location *in* which frame *of a video to minimize semantic class label uncertainty at every pixel, with the smallest computational cost that ensures a given uncertainty bound. We show minimal performance reduction compared to a "paragon" algorithm running all detectors at all locations in all frames, at a small fraction of the computational cost. Our method can handle uncertainty in the labeling mechanism, so it can handle both "oracles" (manual annotation) or noisy detectors (automated annotation).*

1. Introduction

Semantic video segmentation refers to the annotation of each pixel of each frame in a video with a class label. If we are given a *data collection mechanism,* either as an "oracle" or a detector for each known object class, we could perform semantic video segmentation in a brute-force (and simplistic) way by labeling each pixel in each frame. Such a "baseline" algorithm is clearly inefficient as it fails to exploit spatio-temporal regularities in the video signal. Moreover, capturing and exploiting these regularities is computationally inexpensive and can be done using a variety of low-level vision techniques. On the other hand, detecting and localizing objects in the scene requires high-level semantic procedures that have far greater computational cost (in the manual annotation scenario, semantic procedures are replaced with an expensive human annotator). In other words, the complexity of annotating a video sequence is dominated by the cost of high-level procedures, *e.g.* submitting images to a battery of detectors. The annotation cost decreases if fewer such procedures are performed.

We describe a method to reduce the complexity of a labeling scheme, using either an oracle or a battery of detectors,

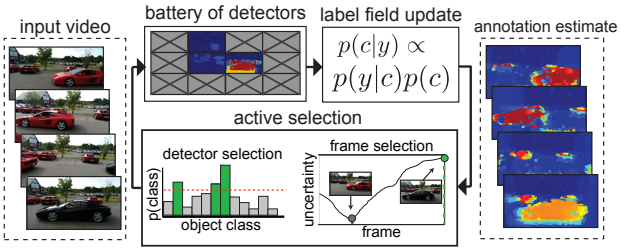

Figure 1. *Given an input video, our approach iteratively improves the annotation estimate by submitting "informative" frames to a "relevant" subset of object detectors. At each iteration, we select the most informative frame (and possibly region within it) based on uncertainty of current annotations. We then select a subset of relevant object detectors, based on estimate of classes that are present in the video. Responses of these detectors are used to update the posterior of the label field, which is then used to perform selection at the next iteration.*

by exploiting temporal consistency and actively selecting *which* data to gather (which detector), *when* (which frame) and *where* (location in an image).

It is important to stress that our method aims to *reduce complexity*, but in principle *can do no better than the baseline*, since it is using only a subset of the data. To avoid confusion, we call the performance upper bound *paragon*, rather than baseline. If the data collection mechanism is reliable (*e.g.* an oracle), we show minimal performance reduction at a fraction of the cost.

Our approach is framed as *uncertainty reduction* with respect to the choice of frame, detector, and location. As a result, we can work with uncertain data collection mechanisms, unlike many label propagation schemes that assume an oracle [31]. As output, we provide a class-label probability distribution per pixel, which can be used to estimate the most likely class, and also provides the labeling uncertainty.

Our method hinges on the *causal* decision of what *future* data to gather, when, and where, based on inference from past data, so as to reduce labeling uncertainty (or "informa-

tion gain" [19]). The method is formulated as a stochastic subset selection problem, finding an optimal solution to which is intractable in general. However, the problem enjoys the *submodular* property [22], so a greedy heuristic attains a constant factor approximation of the optimum [17], a fact that we exploit in our method. A brief overview of our framework is shown in Fig. 1.

1.1. Related work and contributions

We are motivated by the aim to perform semantic video segmentation using as few resources (frames and detectors) as possible, while guaranteeing an upper bound on residual uncertainty. We do so by sequentially choosing the best measurements, which relates to *active learning*. Searching for the best region within the image relates to *location selection*. Detector selection is performed by leveraging object co-occurrences in the video; thus, work on *contextual information* is relevant.

Active learning aims to minimize complexity by selecting the data that would provide the largest "value" relative to the task at hand. It has been used in image segmentation [28] and user-guided object detection [33]; tradeoffs between cost and informativeness were studied in [30], and a procedure efficiently computing the objective was described in [21]. The active learning framework has been used for keyframe selection in an interactive video annotation task in [32]. In a work closest to ours, [31] addresses frame selection for label propagation in video. However, their method relies on oracle labeling and moreover cannot be easily extended to location selection.

Location selection has been studied to reduce the number of evaluations of a sliding-window detector. Previously this was done by generating a set of diverse image partitions likely to cover an object [1]. In [27], it was shown that using such segmentation-driven approach on large image databases maintains performance, while providing computational savings. In [2] a class-dependent sequential decision ("where to look next") approach exploits "context" learned in training, but is limited to finding a single object per image. Recently, a sequential decision strategy that requires user interaction, but does not have this limitation was described in [6]. Our approach is not limited to a single object, is class-independent, and is based on direct uncertainty-minimization framework.

Contextual information has been used to prune false positives in *single images* [7] by using co-occurrence statistics and learning dependencies among object categories. Similarly, [34] use a conditional random field to infer "category presence" using co-occurrence statistics in single images. Our work is related to [16], who exploit co-occurrences to sequentially choose which *detectors* to run to improve speed of object recognition in single images. On the other hand, we tackle video, which allows us to obtain significant computational savings by not running "unnecessary" detectors.

Figure 2. *A pair of frames with temporally consistent regions ($\{S_i\}$, [5]). The highlighted regions are present in both frames.*

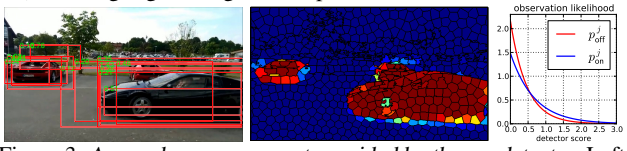

Figure 3. *A pseudo-measurement provided by the car detector.* Left: *a set of bounding boxes given by a DPM detector [10].* Middle: *segmentation result using Grabcut [23]. Color indicates detector score. This is taken as the measurement in our framework.* Right: *likelihood (1) for a "car" class.*

In this paper, we focus only on labeling "objects" in video, as found by bounding box detectors. Extending the approach to also use region-based detectors, as in [26], is possible, and could allow for using geometric context [12].

Our *first contribution* is a frame selection method for video annotation, which naturally allows for uncertainty in the measurements, and thus is applicable to both a battery of standard object detectors (yielding automated annotation), as well as an error-free oracle (yielding manual annotation). Region selection within an image is then naturally added to the framework – this is our *second contribution*. Our *third contribution* is the extension of the framework to enable not just the frame and location selection, but also the selection of detectors based on video shot context.

2. Formulation

In Sec. 2.1 we give an overview of our probability model, introduce object detectors, and describe how their outputs are used to infer the underlying object class labels. Sec. 2.2 introduces the information gathering framework, and proposes an online strategy to select the most informative frames on which to run detectors. This strategy is extended to selecting a region within an image in Sec. 2.3. Sec. 2.4 describes a method for selecting the best subset of detectors by inferring and exploiting context in the video.

Let $I(t) : D \to \mathbb{Z}_+$ be the image defined on a domain $D \subset \mathbb{R}^2$, and let $\{I(t)\}_{t=1}^F$ be F frames of a video. The set of temporally consistent regions (*e.g.* supervoxels or other oversegmentation of the video) $\{S_i\}_{i=1}^N$ forms a partition of the video domain $D^F = \cup_{i=1}^N S_i$ with $S_i \cap S_j = \varnothing$ (see Fig. 2). We assume that these regions respect object boundaries, so that it is possible to associate (temporally-consistent) object class labels to them. For the i-th region, we denote such label by $c_i \in \{0, \ldots, L\}$.

A bank of object detectors represents a set of "test functions" that, when executed at time t, provide a measurement

$y(t) : D \to \mathbb{R}_+^L$. We are interested in labels assigned to regions, so we will write $y_i(t) \in \mathbb{R}_+^L$ to denote responses of a detector bank supported on subset of S_i in t-th frame. Its j-th component $y_i^j(t) \in \mathbb{R}_+$ is the "detection score" for object class $j \in \{1, \dots, L\}$[1] (see Fig. 3) that provides uncertain evidence for the underlying labels.

2.1. Probability model

To measure the "value" of future measurements y for the task of inferring labels $\{c_i\}_{i=1}^N$, we need to quantify their uncertainty before actually measuring them, which requires a probabilistic detector response model. We assume object labels to be spatially independent: $p(c_1, \dots, c_N) = \prod_{i=1}^N p(c_i)$, with a uniform prior $p(c_i) = \frac{1}{L+1}$. This assumption is introduced for simplicity, and can be lifted by using the Markov random field (MRF) model.

Detector response model. When a bank of detectors is deployed on t-th frame, we obtain evidence for the object labels of regions supported in that frame. For i-th region, this evidence is modeled by a likelihood $p(y(t)|c_i) = p(y_i(t)|c_i)$ (responses whose domain does not intersect the i-th region are not informative). Moreover, we assume that individual detector responses are conditionally independent given the label: $p(y_i(t)|c_i) = \prod_{j=1}^L p(y_i^j(t)|c_i)$. To learn these distributions, we use VOC 2010 database. Namely, for each detector, we learn the "true positive" $p_{j,\mathsf{on}}(y^j) \doteq p\big(y^j | c = j\big)$ and the "false positive" $p_{j,\mathsf{off}}(y^j) \doteq p\big(y^j | c \neq j\big)$ distributions, which we model as exponentials:

$$\begin{cases} p_{j,\mathsf{on}}(y^j) & = \lambda_{j,\mathsf{on}} \exp(-y^j \lambda_{j,\mathsf{on}}) \\ p_{j,\mathsf{off}}(y^j) & = \lambda_{j,\mathsf{off}} \exp(-y^j \lambda_{j,\mathsf{off}}) \end{cases} \tag{1}$$

As shown in Fig. 3, $p_{j,\mathsf{off}}$ decays faster than $p_{j,\mathsf{on}}$ – false positive responses usually have smaller scores than true positive responses. Using these distributions, for a background class ($c_i = 0$), we have $p(y_i(t)|c_i = 0) = \prod_{j=1}^L p_{j,\mathsf{off}}(y_i^j(t))$, while for an object class ($k \geq 1$), we have $p(y_i(t)|c_i = k) = p_{k,\mathsf{on}}(y_i^k(t)) \prod_{\substack{j=1 \\ j \neq k}}^L p_{j,\mathsf{off}}(y_i^j(t))$.

Oracle response model. An error-free oracle can be viewed as an ideal detector that returns "1" if an object of a particular class is present at region i and "0" otherwise. In this case, the (binary) measurements $y_i^j(t) \in \{0, 1\}$ are deterministic functions of the underlying labels, and can be written using Kronecker deltas as: $y_i^j(t) = \delta(j - c_i)$, with the likelihood:

$$p(y_i^j(t)|c_i) = \begin{cases} 1 - \epsilon & \text{if } y_i^j = 1 \\ \epsilon & \text{if } y_i^j = 0 \end{cases}. \tag{2}$$

The regularization by ϵ (a small number) is needed to avoid singularities due to violations of modeling assumptions (either oracle errors, or failure of labels' temporal consistency). Because our goal is automated annotation, we will refer to detector responses throughout the paper; however, the methods can be directly applied to oracle annotation as well.

[1] We have $L+1$ object classes and L detectors, since there is no standard "background detector".

Label field update. Given detector responses in frame t, the label probability in region i is updated as

$$p\big(c_i | y(t)\big) \propto p\big(y_i(t) | c_i\big) p\big(c_i\big). \tag{3}$$

This can be extended to a recursive update: if $\mathcal{Y}^k = \big\{y(t_1), \dots, y(t_k)\big\}$ is the history of detector responses taken at frames $\{t_j\}_{j=1}^k$, then

$$p\big(c_i | \mathcal{Y}^k\big) \propto p\big(y_i(t_k) | c_i\big) p\big(c_i | \mathcal{Y}^{k-1}\big), \tag{4}$$

which is standard in recursive Bayesian filtering [14]. When $k = 1$ (the first update) it simply restates (3) (Bayes' rule).

2.2. Information gathering; active frame selection

Having described the update of label probabilities after measuring detector responses, we return to the question of selecting the best subset of frames where to run them. Our *goal* is to maximize the "Value of Information", (VoI [11, 13]), or *uncertainty reduction*, with respect to the *choice* of frames to submit to the oracle, or to test against a battery of detectors. Thus, given frames $\{I(t)\}_{t=1}^F$, regions $\{S_i\}_{i=1}^N$, and a budget of K frames, we minimize uncertainty on labels $\{c_i\}_{i=1}^N$ with respect to a selection of K measurements (yet) to be taken

$$t_1^*, \dots, t_K^* = \operatorname*{argmin}_{\mathcal{T}:|\mathcal{T}| \leq K} H\big(c_1, \dots, c_N \big| y(t_1), \dots, y(t_K)\big). \tag{5}$$

We measure uncertainty by (Shannon) entropy [8].

The problem (5) is in general intractable. For a very special case of noise-free measurements, a dynamic programming (DP) solution exists [18, Alg.1], as was also shown in [31]. When measurements are noisy, this solution is not guaranteed to be optimal. However, due to conditional assumptions made in Sec. 2.1, the problem is *submodular*, so a greedy decision policy yields a constant factor approximation to the optimum [17]. Thus we settle for

$$s^* = \operatorname*{argmin}_s H\big(c_1, \dots, c_N \big| y(s), \mathcal{Y}^k\big), \tag{6}$$

where $\mathcal{Y}^k = \{y(t_1), \dots, y(t_k)\}$ is the set of *already observed* detector responses, and s is the frame index for the next measurement *yet to be taken*. This policy begins with $\mathcal{T} = \varnothing$, at each stage chooses the frame s^* that provides the greatest uncertainty reduction, updates the set of chosen frames $\mathcal{T} := \mathcal{T} \cup \{s^*\}$, and repeats. Notice that this policy has two advantages over the DP solution: first, it is *online* and thus does not require a pre-defined budget (K). Second, it is less susceptible to modeling errors because it uses *observed* responses in making the next decision.

Using the properties of conditional entropy we can write $H(c_1,\dots,c_N|y(s),\mathcal{Y}^k) = H(c_1,\dots,c_N|\mathcal{Y}^k) - \mathbb{I}(y(s);c_1,\dots,c_N|\mathcal{Y}^k)$. Since the first term (uncertainty of c_i's *before* the next selection) is independent of $y(s)$, (6) is equivalent to maximizing the second term, which is the mutual information (MI) between the *next* measurement and the labels: $\mathbb{I}(y(s);c_1,\dots,c_N|\mathcal{Y}^k)$.

Due to the spatial independence of labels, we have that $\mathbb{I}(y(s);c_1,\dots,c_N|\mathcal{Y}^k) = \sum_{i=1}^N \mathbb{I}(y(s);c_i|\mathcal{Y}^k)$. Thus, we can rewrite (6) as

$$s^* = \arg\max_s \sum_{i=1}^N \mathbb{I}(y(s);c_i|\mathcal{Y}^k) \tag{7}$$

Moreover, if c_i is not present in frame s, then $y(s)$ does not provide evidence for it, and $\mathbb{I}(y(s);c_i|\mathcal{Y}^k)=0$. On the other hand, if c_i is present, $y(s)$ is informative – proportionally to uncertainty in c_i. Thus, the criterion prefers frames that have the largest number of uncertain regions. As in the twenty-question game, we wish to label the data that is *most uncertain* given prior measurements. Taking measurements on frames that have little uncertainty provides little information gain.

There are no closed form expression for mutual information for the densities that we consider here. Hence, we need to either approximate it by Monte Carlo sampling, or to find efficiently computable proxies. Because we are interested in a maximization problem, the natural proxy of interest is the lower bound on $\mathbb{I}(y(s);c_i|\mathcal{Y}^k)$. However, it is also very common to use upper bounds (most often using the second-moment Gaussian approximation). Upper bounds are acceptable because we are ultimately interested in the maximizing *point*, rather than the maximizing *value*. Thus, the tightness of the bound is irrelevant, and it is only required that the maxima are preserved. We use an upper bound:

$$\mathbb{I}(y(s);c_i|\mathcal{Y}^k) \le \sum_{m=0,n=0}^{L,L} w_m w_n \sum_{j=1}^{L} \frac{\eta_{jn}}{\eta_{jm}} - L \tag{8}$$

where $w_m \doteq p(c_i=m|\mathcal{Y}^k)$, and $\eta_{jm}=\lambda_{j,\mathsf{on}}$ if $j=m$ and $\eta_{jm}=\lambda_{j,\mathsf{off}}$ otherwise. We prove this result in the technical report [15] and empirically show that it preserves the local maxima of the Monte Carlo approximation.

As an alternative to (7) we can try to select frames with high classification error. Let $\mathcal{E}(s,i) = \mathsf{alive}(S_i,s)\big(1-\max_\ell p(c_i=\ell|\mathcal{Y}^k)\big)$ where $\mathsf{alive}(S_i,s)=1$ if region S_i is supported on frame s and is 0 else, and the second term is simply the classification error. We then use a criterion

$$s^* = \arg\max_s \sum_{i=1}^N \mathcal{E}(s,i). \tag{9}$$

Note that unlike MI, this criterion does not make predictions about the next measurement $y(s)$, and is therefore very simple to compute. Yet, as we show in Sec. 3, it performs competitively in practice.

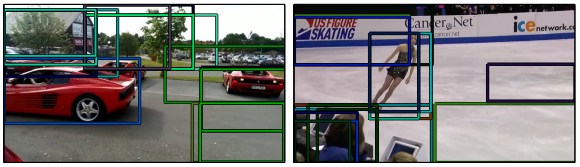

Figure 4. *Candidate regions used for selecting the most informative location* (10)

2.3. Active location selection

It is straightforward to augment the frame selection criterion (7) with the selection of the most informative region R ($R \subseteq D$) within the image (to either run detectors on, or to query an oracle). In this case, at each stage we maximize $\sum_{i=1}^N \mathbf{1}_{\{S_i \subset R\}} \mathbb{I}(c_i;y(s)|\mathcal{Y}^k)$ over a pair (s,R), where the indicator $\mathbf{1}_{\{S_i\subset R\}}$ discards all the labels c_i that are outside R (the region that is submitted to detectors). However, because the mutual information is nonnegative, the best region chosen by this strategy will always contain the entire image ($R^* = D$), and a proper subset of the image domain will never be chosen. Thus, it is necessary to associate a *cost* to R. We choose a cost that is proportional to the region size, and trade off the two as:

$$s^*, R^* = \arg\max_{s,R} \sum_{i=1}^N \mathbf{1}_{\{S_i\subset R\}} \mathbb{I}(c_i;y(s)|\mathcal{Y}^k) - \gamma|R| \tag{10}$$

with γ – a weighing term. A more sophisticated approach could estimate and use the computational effort in running a detector bank as a function of $|R|$ (which needs not be linear). In practice, we maximize the criterion over a finite, diverse set of candidate regions (shown in Fig. 4), which presumably cover objects in the image.

2.4. Context and active detector selection

In situations where L (the number of available detectors) is large, but the number of classes present in the scene is small, it is not computationally efficient to run the entire battery of detectors. To address this issue, we extend the framework to include not just a selection of subsets of frames and regions to be labeled, but also a selection of the subset of detectors to be deployed, by exploiting context in the video.

Probability model. To describe the *context* of the video sequence, we introduce a random variable $o=(o_1,\dots,o_L)\in\{0,1\}^L$ that is global within a video shot, where o_j represents presence or absence of k-th object category in the shot. Detector responses provide soft evidence as to whether object categories are present in the shot. Our belief in objects being present is summarized by the distribution $p(o|\mathcal{Y}^k)$ – the posterior of the context variable, given evidence from the detectors.

To infer this distribution, we must first specify the likelihood $p(y|o)$. We assume that the distribution can be factorized as $p(y|o) = \prod_{j=1}^{L} p(y^j|o_j)$, where each term is a model for a detector response given that an object j is present (or absent). To be invariant to response location, we use the maximum detector response score within an image $z^j(t) = \max_i y_i^j(t)$ as the observation associated with j-th category presence, and specify the model as:

$$\begin{cases} p(y^j(t)|o_j = 0) & = p_{j,\text{off}}\big(z^j(t)\big) \\ p(y^j(t)|o_j = 1) & = \pi p_{j,\text{on}}\big(z^j(t)\big) + (1-\pi)p_{j,\text{off}}\big(z^j(t)\big). \end{cases} \tag{11}$$

where $p_{j,\text{off}}$, $p_{j,\text{on}}$ are the distributions used in (1). The density $p(y^j(t)|o_j = 1)$ is a mixture, and can account for the possibility of an object *not* being present in frame t despite being present in the video shot. The mixture parameter π is related to the fraction of time the object is expected to be visible in the shot.

Detector selection. The marginal distributions $p(o_j|\mathcal{Y}^k)$ describe the probability that j-th class is present in the video, given the observation history. If computation is limited, when $p(o_j|\mathcal{Y}^k)$ is small, we should avoid running j-th detector. This can be phrased in terms of a threshold α on marginal probabilities, yielding a two-stage procedure:

$$\begin{cases} J & = \{j : p(o_j|\mathcal{Y}^k) > \alpha\} \\ s^* & = \arg\max_s \sum_{i=1}^{N} \mathbb{I}(\{y^j(s)\}_{j\in J}; c_i|\mathcal{Y}^k) \end{cases} \tag{12}$$

where $\{y^j(s)\}_{j\in J}$ is a set of responses for detectors indexed by J. This procedure is performed at each stage of our sequential decision problem: once the set J of object categories is chosen, we select the most informative frame s^* to run these detectors on, acquire new evidence, update the posteriors, and repeat.

Of course, the frame selection step in the detector selection procedure (12) can be extended to allow for region selection. Due to space constraints, we do not explicitly write out the equation; however, it is no different from the extension of the original frame selection criterion (7) to *frame and region* selection (10).

3. Experiments

To evaluate our approach, we use public benchmark datasets including human-assisted motion[20], MOSEG[4], Visor[29], and BVSD[25], as well several videos from Flickr (Fig. 5). We are interested in classification error, so we manually created pixelwise ground truth by labeling these sequences.

Implementation Details. We use [5] to compute video-superpixels (temporally consistent regions), with 500-700 superpixels in an image. Our detectors contain models for 20 classes, pre-trained on VOC 2010 [9], based on DPM [10], and refined with GrabCut [23], as shown in Fig. 3. We offset detector scores to make them nonnegative and convert them into likelihoods for use in our model (1), with likelihood distribution learned from the VOC 2010. To approximate $p(o|\mathcal{Y}^k)$, we use a fully connected MRF, with node and edge potentials learned using the UGM toolbox [24]. The co-occurrence statistics are derived from VOC. Throughout all experiments we used $\alpha = 0.005$ (detector selection threshold (12)), $\pi = 0.5$ (mixture weight in (11)). As written in (7) and (10), the selection criteria weigh terms corresponding to different regions equally. In practice we weighted the terms according to region size; this choice improved performance.

Figure 5. Sample frames from a subset of video sequences used for testing our algorithm, from BVSD ([25],top) and Flickr (ours, bottom).

Videos in our database vary in duration (from 19 to 300 frames), so we report classification accuracy as a function of percentage of sequence labeled; this can be either a percentage of *frames* submitted to detectors, or a percentage of *pixels* in a video submitted to detectors.

Frame selection. Our first experiment compares our information-seeking approach (with criteria given by (8) and (9)) with the naive methods (uniform and random selection). We also compare with the DP approach of [31]: we use their selection criterion, but propagate labels using our model, to make the comparison with the other methods fair. As can be seen in Table 1(top), we consistenly outperform other methods when the error-free "oracle" is used. The improvement over [31] is due to our objective being closely related to the labeling error and birth/death of temporal regions, whereas their selection involves a combination of optical flow, image intensities, and backward-forward optical flow consistency (a proxy for occlusions).

When the "oracle" is replaced by a battery of detectors, the DP approach is not optimal. Moreover, due to erroneous detections (false positives or misses), *any* of the methods is susceptible to failure, even if they choose "informative" frames. We observe this in Table 1 (bottom): although our methods perform best on average, at 10% labeling, "uniform" attains a lower classification error. When the detector performance is poor, any sampling scheme yields equally bad or worse performance.

Location selection. Our second experiment compares our region and frame selection scheme against the random selection. The approach in [31] cannot be easily extended to region selection, so we do not compare against it. Our

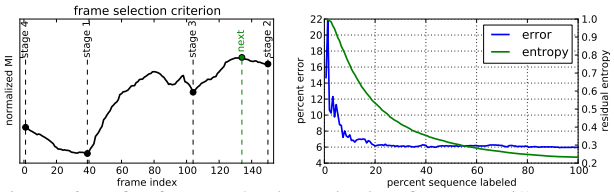

Figure 6. Left: *frame selection criterion for "Ferrari" sequence (marked with "*" in Fig. 5), for the first few selection stages. The selected frames are not uniformly distributed.* Right: *classification error and normalized residual entropy ($\sum_{i=1}^{N} H(c_i|\mathcal{Y}^k)$) as a function of number of frames selected for labeling. Note that after 20% sequence is labeled, error reaches an asymptote.*

%	$\mathcal{E}$ (our)	MI (our)	[31]	uniform	random
1	**4.666**	**4.666**	5.112	6.079	5.304 ± 0.29
5	**2.696**	2.700	3.130	3.270	3.556 ± 0.19
10	**2.264**	2.274	2.434	2.478	2.912 ± 0.14
15	**2.088**	2.092	2.110	2.194	2.586 ± 0.10
20	**2.005**	2.007	2.023	2.072	2.455 ± 0.06
30	1.935	**1.930**	1.961	1.979	2.358 ± 0.07

%	$\mathcal{E}$ (our)	MI (our)	[31]	uniform	random
1	**11.224**	11.277	12.623	11.948	12.624 ± 1.0
5	**9.069**	9.129	10.446	9.363	9.356 ± 0.43
10	8.097	8.393	8.455	**7.764**	8.639 ± 0.44
15	7.862	**7.674**	8.389	7.844	8.682 ± 0.37
20	7.611	**7.582**	8.041	7.589	8.499 ± 0.26
30	7.449	**7.377**	8.105	7.890	8.221 ± 0.25

Table 1. *Average classification error of different frame selection strategies with oracle (top) and detector (bottom) labeling; 1%-30% frames.*

candidate regions vary in size and location, and uniformly selecting representatives out of this set is rather problematic; therefore we do not test against this approach.

We compute candidate regions using [27], which typically consist of bounding boxes that entirely contain objects of interest (see Fig. 4). Typically, per image, we have 10-20 regions that occupy 10%-80% of the image. Results with oracle and detector labeling are shown in Table 2, as a function of percent pixels used to obtain a labeling (proportional to the sum of selected regions' areas). Perhaps unsurprisingly, the "random" selection performs poorly.

Detector selection. This experiment demonstrates the possibility of reducing computation effort without suffering a performance penalty, by reducing the number of detectors deployed at each stage. In these experiments we perform frame selection using the MI criterion (8) (although $\mathcal{E}$ can be used as well). We do not perform region selection. The typical behavior of the detector selection, as shown in Fig. 7, is to run fewer detectors as more and more frames are selected. Often, in the limit, only the detectors for the classes that are present are fired. Thus, the cost of measurement decreases (and computational savings increase) with number of labeled frames.

One may wonder how much is gained from using co-

%	$\mathcal{E}$(our)	MI(our)	random
1	6.503	**6.189**	9.177 ± 0.502
5	3.434	**3.125**	4.662 ± 0.374
10	2.715	**2.456**	3.273 ± 0.155
20	2.392	**2.187**	2.600 ± 0.067
30	2.289	**2.144**	2.329 ± 0.061

%	$\mathcal{E}$ (our)	MI (our)	random
1	16.878	**16.861**	17.113 ± 1.225
5	**15.266**	15.923	16.061 ± 0.458
10	14.588	**13.830**	15.192 ± 0.716
20	12.987	**12.372**	13.935 ± 0.304
30	11.970	**11.819**	12.963 ± 0.182

Table 2. *Average classification percent error of different region+frame selection strategies with oracle (top) and detector (bottom) labeling; using 1%-30% pixels.*

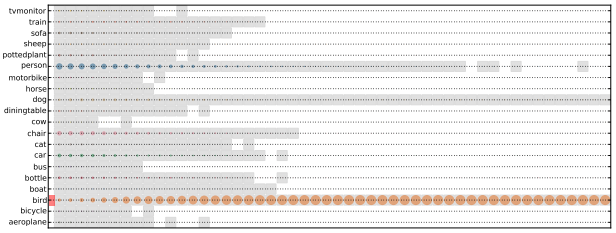

Figure 7. *Detector selection on "Planet Earth" sequence (marked with "#" in Fig. 5): For each frame selected for labeling (abscissa), deployed detectors are shown as gray boxes. Colored circles represent $p(o_j|\mathcal{Y}^k)$ – the belief of a particular class being present, with the area proportional to the value of the posterior, and ground truth is indicated on the left by a red mark (bird). The "dog" class is fired the longest because it co-occurs with "bird" in the training set.*

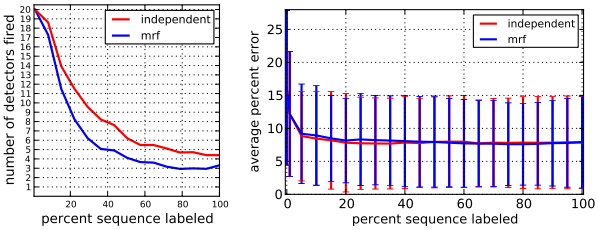

Figure 8. Left: *As more frames get labeled, fewer detectors are fired. We show the average number of detectors fired, over all sequences in our datasets, for "MRF" and "independent" approximations. MRF approximation of the joint distribution makes it possible to quickly stop firing contextually irrelevant detectors.* Right: *classification error as function of labeled frames. The slightly larger error in "MRF" is the price paid for reduced number of used detectors.*

occurrence information. To investigate this, we performed a set of experiments under the independence assumption $p(o|\mathcal{Y}^k) = \prod_{j=1}^{L} p(o_j|\mathcal{Y}^k)$. The average computation savings and the average classification errors are shown in Fig. 8. Using an MRF, we get a substantial decrease in the number of detectors that are fired at each stage. This is because co-occurrence information allows us to quickly suppress probabilities for contextually atypical situations.

Baseline and "paragon" annotation. We first illustrate the gain from using temporal information. We compare our

quantity	method	cost
$y(t)$	[10]+[23]	$60s + 180s$
$y(t)\mathbf{1}_R$	[10]+[23]	$\frac{\lvert R\rvert}{\lvert D\rvert}(60s + 180s)$
$\{y^j(t)\}_{j\in J}$	[10]+[23]	$\frac{\lvert J\rvert}{L}(60s + 180s)$
$\{S_i\}_{i=1}^N$	[3]+[5]	$F(10s + 5s)$
candidate regions	[27]	$F(0.5s)$
frame selection	(7) + (4)	$0.02s$
frame+region selection	(10) + (4)	$0.045s$
detector selection	(12)	$5s$

Table 3. A summary of quantities that are computed in our framework, associated procedures, and costs, measured in terms of time.

approach with the one that does not use temporal consistent regions. Specifically, our "probabilistic baseline" (PB) is the maximum likelihood estimation using the model (1), applied to the detected and subsequently segmented regions. It considers detector responses from *all* frames, but treats each frame independently. Our framework outperforms this approach; in fact, we perform better after labeling only a small fraction of the sequence. Fig. 9 shows the percentage of frames needed to reach the performance of this baseline: we perform better after labeling only 10% of the sequence. Fig. 10 shows several examples of annotation using PB and our approach: temporal regularities allow us to suppress a large number of false positive detections.

We also compare against the "paragon" approach, which uses *all* frames, *all* detectors, and temporal consistency. But by running detectors on only a fraction of the frames, we do not perform significantly worse. As shown in Fig. 11, classification error has a "diminishing returns" property: as more frames are labeled, the improvement is decreasing. This suggests that using *all* frames is unnecessary, and if one has computational constraints, "early stopping" can be beneficial.

Computation savings. To produce a PB annotation, one runs detectors and segmentation on every frame. DPM [10] takes 60s/frame (for 20 classes) and Grabcut[23] takes 180s/frame (we segment every bounding box using an unoptimized MATLAB implementation); the sum of the two is the "cost" of observing $y(t)$. To leverage temporal consistency, we use [5] (5s/frame) and optical flow [3] (10s/frame). The costs of our *frame selection* framework are negligible: computation of frame selection utility (7) (or (9)) takes 15 ms/stage on all frames, inference (4) takes 5 ms/stage, both measured on the longest sequence. Region selection requires the candidate regions [27], which cost 0.5s/frame, but computation of location selection utility (10) remains negligible (40 ms/stage). Our *detector selection* framework requires estimating "presence" marginals $p(o_j|\mathcal{Y}^k)$ at a cost of 5s per stage of the algorithm. These "costs" are in Table 3.

We can estimate the PB cost as $F(240s)$, where F is the number of frames in the sequence. The "paragon" requires temporally consistent regions $\{S_i\}_{i=1}^N$, and thus costs $F(270s)$. The frame selection framework requires negligible computation per stage and reduces computation cost to $F(15s) + K(240s)$, where K is the budget of frames to be labeled (according to Fig. 11, $K \approx 0.2F$ is sufficient). The region selection framework decreases the observation cost linearly in region size; an admittedly coarse assumption. The cost of a measurement supported on region R, denoted $y(t)\mathbf{1}_R$, is then reduced from $240s$ to $|R|/|D|(240s)$. The detector selection framework decreases the observation cost to $|J|/L(240s)$, but incurs an additional $5s$ per stage (due to context inference). As a specific example, for a "Ferrari" sequence with $F = 150$, PB costs 600 min. The "paragon" costs 638 min. Our framework with frame selection and 20% labeling needs only ∼158 min. Frame and region selection costs the same amount. Using frame and detector selection framework, we use 54% detectors in the first 20% frames, reducing the cost to just ∼85 min.

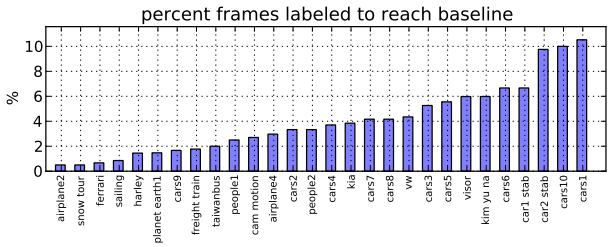

Figure 9. *Percentage and number of frames to be labeled by detectors to match PB performance. On average across all sequences we only need to label* 4.012% *of the frames. Baseline obtains* 12.669% *average error using detectors on all frames.*

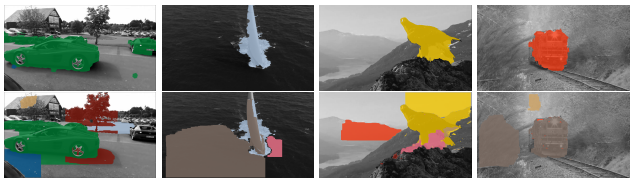

Figure 10. Top*: Sample frames with our annotation using 20% of the sequence.* Bottom*: PB labeling on the same frames. Different colors correspond to different object classes. Temporal regularities allow us to remove many false positive detections present in PB.*

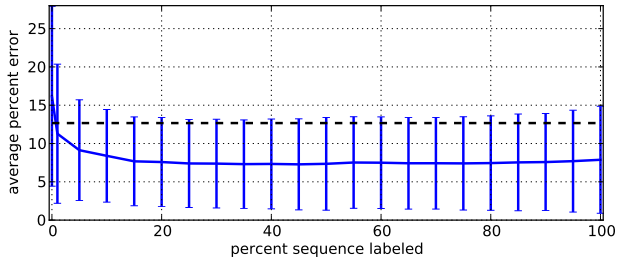

Figure 11. *Blue: average error over entire dataset as a function of labeled frames. As more frames are labeled, the improvement in error decreases (on average), suggesting that if computational budget is limited, it is unnecessary to use all frames. Black dashed line is PB (which uses all frames independently).*

4. Discussion

We have presented an uncertainty-based active selection approach to determine *which locations* of *which frames* in a video shot to run *which detector* on to arrive at a labeling of each pixel at the smallest computational cost that ensures a bound on residual uncertainty ($\sum_i H(c_i|\mathcal{Y}^k)$). We proposed two information-seeking criteria, MI and $\mathcal{E}$, and demonstrated that they outperm other selection schemes.

Unlike existing label propagation schemes that assume an oracle, we can handle uncertainty in the measurements, by leveraging an explicit probabilistic detector response model, a prior on classes learned from the PASCAL VOC dataset, and a hidden context variable global to each video shot. Our method is causal, respects the spatio-temporal regularities in the video, and falls within the class of submodular optimization problems that enjoy desirable bounds in performance of greedy inference relative to the (intractable) optimum.

We compare the performance of our scheme on various baselines, including "paragons" running all detectors at all locations in all frames. In the presence of reliable detectors (an oracle, in the limit), a manyfold reduction of computational cost is possible with negligible performance drop.

Acknowledgments. Supported on AFRL FA8650-11-1-7156:P00004, ARO MURI W911NF-11-1-0391, and ONR N00014-13-1-0563.

References

[1] B. Alexe, T. Deselaers, and V. Ferrari. What is an object? In *CVPR*, 2010. 2

[2] B. Alexe, N. Heess, Y. W. Teh, and V. Ferrari. Searching for objects driven by context. In *NIPS*, December 2012. 2

[3] A. Ayvaci, M. Raptis, and S. Soatto. Sparse occlusion detection with optical flow. *IJCV*, 2012. 7

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

[5] J. Chang, D. Wei, and J. W. F. III. A Video Representation Using Temporal Superpixels. In *CVPR*, 2013. 2, 5, 7

[6] Y. Chen, H. Shioi, C. F. Montesinos, L. P. Koh, S. Wich, and A. Krause. Active detection via adaptive submodularity. In *ICML*, 2014. 2

[7] M. J. Choi, J. Lim, A. Torralba, and A. Willsky. Exploiting hierarchical context on a large database of object categories. In *CVPR*, 2010. 2

[8] T. M. Cover and J. Thomas. *Elements of Information Theory*. Wiley, 1991. 3

[9] M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2010 (VOC2010) Results. 5

[10] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *TPAMI*, 2010. 2, 5, 7

[11] D. Heckerman, E. Horvitz, and B. Middleton. An approximate nonmyopic computation for value of information. *TPAMI*, 1993. 3

[12] G. Heitz and D. Koller. Learning spatial context: Using stuff to find things. In *ECCV*, 2008. 2

[13] R. A. Howard. Information value theory. *IEEE Trans. Systems Science and Cybernetics*, 1966. 3

[14] A. Jazwinski. *Stochastic Processes and Filtering Theory*. Mathematics in science and engineering. 1970. 3

[15] V. Karasev, A. Ravichandran, and S. Soatto. Active frame, location, and detector selection for automated and manual video annotation. In *Tech Report UCLACSD140010*, 2014. 4

[16] S. Karayev, T. Baumgartner, M. Fritz, and T. Darrell. Timely object recognition. In *NIPS*, 2012. 2

[17] A. Krause and C. Guestrin. Near-optimal nonmyopic value of information in graphical models. In *UAI*, 2005. 2, 3

[18] A. Krause and C. Guestrin. Optimal value of information in graphical models. *JAIR*, 2009. 3

[19] D. Lindley. On the measure of the information provided by an experiment. *Annals of Mathematical Statistics*, 1956. 2

[20] C. Liu, W. T. Freeman, E. H. Adelson, and Y. Weiss. Human-assisted motion annotation. In *CVPR*, 2008. 5

[21] O. Mac Aodha, N. D. F. Campbell, J. Kautz, and G. J. Brostow. Hierarchical subquery evaluation for active learning on a graph. In *CVPR*, 2014. 2

[22] G. Nemhauser, L. Wolsey, and M. Fisher. An analysis of approximations for maximizing submodular set functions. *Mathematical Programming*, 1978. 2

[23] C. Rother, V. Kolmogorov, and A. Blake. "Grabcut": interactive foreground extraction using iterated graph cuts. *ACM Trans. Graph.* 2, 5, 7

[24] M. Schmidt. UGM:Matlab toolbox for probabilistic undirected graphical models. www.di.ens.fr/mschmidt/Software/UGM.html. 5

[25] P. Sundberg, T. Brox, M. Maire, P. Arbelaez, and J. Malik. Occlusion boundary detection and figure/ground assignment from optical flow. In *CVPR*, 2011. 5

[26] J. Tighe and S. Lazebnik. Finding things: Image parsing with regions and per-exemplar detectors. In *CVPR*, 2013. 2

[27] K. E. A. van de Sande, J. Uijlings, T. Gevers, and A. Smeulders. Segmentation as Selective Search for Object Recognition. In *ICCV*, 2011. 2, 6, 7

[28] A. Vezhnevets, J. Buhmann, and V. Ferrari. Active learning for semantic segmentation with expected change. In *CVPR*, 2012. 2

[29] R. Vezzani and R. Cucchiara. Video surveillance online repository (ViSOR): an integrated framework. *Multimedia Tools Appl.*, 2010. 5

[30] S. Vijayanarasimhan and K. Grauman. What's it going to cost you?: Predicting effort vs. informativeness for multi-label image annotations. In *CVPR*, 2009. 2

[31] S. Vijayanarasimhan and K. Grauman. Active frame selection for label propagation in videos. In *ECCV*, 2012. 1, 2, 3, 5, 6

[32] C. Vondrick and D. Ramanan. Video Annotation and Tracking with Active Learning. In *NIPS*, 2011. 2

[33] A. Yao, J. Gall, C. Leistner, and L. J. V. Gool. Interactive object detection. In *CVPR*, 2012. 2

[34] J. Yao, S. Fidler, and R. Urtasun. Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation. In *CVPR*, 2012. 2