# Visual Recognition by Counting Instances: A Multi-Instance Cardinality Potential Kernel

Hossein Hajimirsadeghi Wang Yan Arash Vahdat Greg Mori
School of Computing Science, Simon Fraser University, Canada
hosseinh@sfu.ca, wyan@sfu.ca, avahdat@sfu.ca, mori@cs.sfu.ca

## Abstract

*Many visual recognition problems can be approached by counting instances. To determine whether an event is present in a long internet video, one could count how many frames seem to contain the activity. Classifying the activity of a group of people can be done by counting the actions of individual people. Encoding these cardinality relationships can reduce sensitivity to clutter, in the form of irrelevant frames or individuals not involved in a group activity. Learned parameters can encode how many instances tend to occur in a class of interest. To this end, this paper develops a powerful and flexible framework to infer any cardinality relation between latent labels in a multi-instance model. Hard or soft cardinality relations can be encoded to tackle diverse levels of ambiguity. Experiments on tasks such as human activity recognition, video event detection, and video summarization demonstrate the effectiveness of using cardinality relations for improving recognition results.*

## 1. Introduction

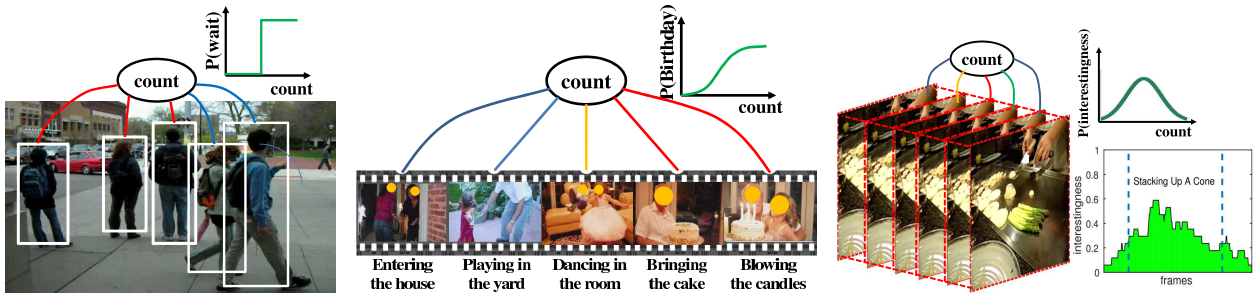
A number of visual recognition problems involve examining a set of instances, such as the people in an image or frames in a video. For example, in group activity recognition (e.g. [6]) the prominent approach to analyzing the activity of a group of people is to look at the actions of individuals in a scene. A number of impressive methods have been developed for modeling the *structure* of a group activity [18, 5, 3], capturing spatio-temporal relations between people in a scene. However, these methods do not directly consider cardinality relations about the *number* of people that should be involved in an activity. These cardinality relations vary per activity. An activity such as a fall in a nursing home [18] is different in composition from an activity such as queuing [5], involving different numbers of people (one person falls, many people queue). Further, clutter, in the form of people in a scene performing unrelated actions, confounds recognition algorithms. In this paper we present a framework built on a latent structured model to encode these cardinality relations and deal with the ambiguity or clutter in the data.

Another example is unconstrained internet video analysis. Detecting events in internet videos [21] or determining whether part of of a video is *interesting* [10] are challenging for many reasons, including temporal clutter – videos often contain frames unrelated to the event of interest or that are difficult to classify. Two broad approaches exist for video analysis, either relying on holistic bag-of-words models or building temporal models of events. Again, successful methods for modeling temporal structure exist (e.g. [8, 28, 25, 29]). Our method builds on these successes, but directly considers cardinality relations, counting how many frames of a video appear to contain a class of interest, and using soft and intuitive constraints such as "the more, the better" to enhance recognition.

Fig. 1 shows an overview of our method. We encode our intuition about these counting relations in a multiple instance learning framework. In multiple instance learning, the input to the algorithm is a set of labeled *bags* containing *instances*, where the instance labels are not given. We approach this problem by modeling the bag with a probabilistic latent structured model. Here, we highlight the major contributions of this paper.

- **Showing the importance of cardinality relations for visual recognition.** We show in different applications that encoding cardinality relations, either hard (e.g. *majority*) or soft (e.g. *the more, the better*), can help to enhance recognition performance and increase robustness against labeling ambiguity.

- **A kernelized framework for classification with cardinality relations.** We use a latent structured model, which can easily encode any type of cardinality constraint on instance labels. A novel kernel is defined on these probabilistic models. We show that our proposed kernel method is effective, principled, and has efficient and exact inference and learning methods.




(a) What is the collective activity? (b) What is this video about? (c) Is this video segment interesting?

Figure 1. Encoding cardinality relations can improve visual recognition. (a) An example of collective activity recognition. Three people are waiting, and two people are walking (passing by in the street). Using only spatial relations, it is hard to infer what the dominant activity is, but encoding the cardinality constraint that the collective activity tends to be the majority action helps to break the tie and favor "waiting" over "walking". (b) A "birthday party" video from the TRECVID MED11 dataset [21]. Some parts of the video are irrelevant to birthdays and some parts share similarity with other events such as "wedding". However, encoding soft cardinality constraints such as "the more relevant parts, the more confident decision", can enhance event detection. (c) A video from the SumMe summarization dataset [10]. The left image shows an important segment, where the chef is stacking up a cone. The right image shows the human-judged interesting-ness score of each frame. Even based on human judgment, not all parts of an important segment are equally interesting. Due to uncertainty in labeling the start and end of a segment, the cardinality potential might be non-monotonic.

## 2. Related Work

This paper presents a novel model for cardinality relations in visual recognition, in particular for the analysis of video sequences. Existing video analysis methods generally focus on structured spatio-temporal models, complementary to our proposed approach. For instance, pioneering work was done by Gupta et al. [8] in analyzing structured videos by creating "storyline" models populated from AND-OR graph representations. Related models have proven effective at analyzing scenes of human activity more broadly in work by Amer et al. [3]. A series of recent papers has focused on the problem of group activity recognition, inferring an activity that is performed by a set of people in a scene. Choi et al. [6, 5], Lan et al. [18], and Khamis et al. [13] devised models for spatial and temporal relations between the individuals involved in a putative interaction. Zhu et al. [33] consider contextual relations between humans and objects in a scene to detect interactions of interest. The structural relations exploited by these methods are a key component of activity understanding, but present different information from the cardinality relations we study.

Analogous approaches have been studied for "unconstrained" internet video analysis. Methods to capture the temporal structure of high-level events need to be robust to the presence of irrelevant frames. Successful models include Tian et al. [28] and Niebles et al. [20], who extend latent variable models in the temporal domain. Tang et al. [25] develop hidden Markov models with variable duration states to account for the temporal length of action segments. Vahdat et al. [29] compose a test video with a set of kernel matches to training videos. Tang et al. [26] effectively combine informative subsets of features extracted from videos to improve event detection. Bojanowski et al. [4] label videos with sequences of low-level actions. Pirsiavash and Ramanan [22] develop stochastic grammars for understanding structured events. Xu et al. [31] propose a feature fusion method based on utilizing related exemplars for event detection. Lai et al. [17] apply multiple instance learning to video event detection by representing a video as multi-granular temporal video segments. Our work is similar in spirit, but contributes richer cardinality relations and more powerful kernel representations; empirically we show these can deliver superior performance.

The continued increase in the amount of video content available has rendered the summarization of unconstrained internet videos an important task. Kim et al. [15] build structured storyline-type representations for the events in a day. Khosla et al. [14] use web images as a prior for selecting good summaries of internet videos. Popatov et al. [23] learn the important components of videos of high-level events. Gygli et al. [10] propose a benchmark dataset for measuring interesting-ness of video clips and explore a set of high-level semantic features along with superframe segmentation for detecting interesting video clips. We demonstrate that our cardinality-based methods can be effective for this task as well, scoring a clip by the number of interesting frames it contains.

### 2.1. Multi-Instance Learning

We develop an algorithm based on multiple instance learning, where an input example consists of a bag of instances, such as a video represented as a bag of frames. The traditional assumption is that a bag is positive if it contains at least one positive instance, while in a negative bag all

the instances are negative. However, this is a very weak assumption, and recent work has developed advanced algorithms with different assumptions [19, 12, 11, 17].

For example, Li et al. [19] formulated a prior on the number of positive instances in a bag, and used an iterative cutting plane algorithm with heuristics to approximate the resultant learning problem. Yu et al. [32] proposed $\propto$SVM for learning from instance proportions, and showed promising results on video event recognition [17]. Our work improves on this approach by permitting more general cardinality relations with an efficient and exact training scheme.

Our approach models a bag of instances with a probabilistic model with a cardinality-based clique potential between the instance labels. This cardinality potential facilitates defining any cardinality relations between the instance labels and efficient and exact solutions for both maximum a posteriori (MAP) and sum-product inference [9, 27]. For example, Hajimirsadeghi et al. [11] used cardinality-based models to embed different ratio-based multiple instance assumptions. Here we extend these lines of work by developing a novel kernel-based learning algorithm that enhances classification performance.

Kernel methods for multiple instance learning include Gärtner et al.'s [7] MI-Kernel, which is obtained by summing up the instance kernels between all instance pairs of two bags. Hence, all instances of a bag contribute to bag classification equally, although they are not equally important in practice. To alleviate this problem, Kwok and Cheung [16] proposed marginalized MI-Kernel. This kernel specifies the importance of an instance pair of two bags according to the consistency of their probabilistic instance labels. In our work, we also use the idea of marginalizing joint kernels, but we propose a unified framework to combine instance label inference and bag classification within a probabilistic graph-structured kernel.

## 3. Proposed Method: Cardinality Kernel

We propose a novel kernel for modeling cardinality relations, counting instance labels in a bag – for example the number of people in a scene who are performing an action. We start with a high-level overview of the method, following the depiction in Fig. 2.

The method operates in a multiple instance setting, where the input is bags of instances, and the task is to label each bag. For concreteness, Fig. 2(a) shows video event detection. Each video is a bag comprised of individual frames. The goal is to label a video according to whether a high-level event of interest is occurring in the video or not. Temporal clutter, in the form of irrelevant frames, is a challenge. Some frames may be directly related to the event of interest, while others are not.

Fig. 2(b) shows a probabilistic model defined over each video. Each frame of a video can be labeled as containing the event of interest, or not. Ambiguity in this labeling is pervasive, since the low-level features defined on a frame are generally insufficient to make a clear decision about a high-level event label. The probabilistic model handles this ambiguity and a counting of frames – parameters encode the appearance of low-level features and the intuition that more frames relevant to the event of interest makes it more likely that the video as a whole should be given the event label.

A kernel is defined over these bags, shown in Fig. 2(c). Kernels compute a similarity between any two videos. In our case, this similarity is based on having similar cardinality relations, such as two videos having similar counts of frames containing an event of interest. Finally, this kernel can be used in any kernel method, such as an SVM for classification, Fig. 2(d).

### 3.1. Cardinality Model

A cardinality potential is defined in terms of counts of variables which take some particular values. For example, with binary variables, it is defined in terms of the number of positively and negatively labeled variables. Given a set of binary random variables $\mathbf{y} = \{y_1, y_2, \cdots, y_m\}$ ($y_i \in \{0, 1\}$), the cardinality potential model is described by the joint probability

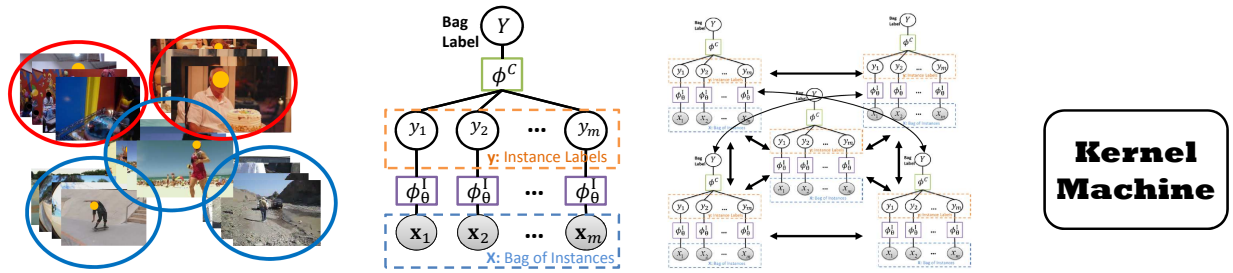
$$P(\mathbf{y}) = \frac{C(\sum_i y_i) \prod_i \exp(\varphi_i y_i)}{\sum_{\mathbf{y}} C(\sum_i y_i) \prod_i \exp(\varphi_i y_i)}, \tag{1}$$

which consists of one cardinality potential $C(\cdot)$ over all the variables and unary potentials $\exp(\varphi_i y_i)$ on features $\varphi_i$ on each single variable. Maximum a posteriori (MAP) inference of this model is straight-forward and takes $O(m \log m)$ time [9]. Sum-product inference is more involved, but efficient algorithms exist [27], computing all marginal probabilities of this model in $O(m \log^2 m)$ time.

In problems with multiple instances, there are assumptions or constraints which are defined on the counts of instance labels. For example, the standard multi-instance assumption states that at least one instance in a positive bag is positive. So, it is intuitive that these constraints can be modeled by a cardinality potential over the instance labels. This modeling helps to have exact and efficient solutions for MIL problems, using existing state-of-the-art inference and learning algorithms.

Using this cardinality potential model as the core, a probabilistic model of the likelihood of a bag of instances $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_m\}$ with the bag label $Y \in \{-1, +1\}$ and the instance labels $\mathbf{y}$ with model parameters $\boldsymbol{\theta}$, is built (c.f. [27]):

$$P(Y, \mathbf{y} | \mathbf{X}; \boldsymbol{\theta}) \propto \phi^C(Y, \mathbf{y}) \prod_i \phi^I_{\boldsymbol{\theta}}(\mathbf{x}_i, y_i). \tag{2}$$




(a) Preparing Training Data as Positive and Negative Bags of Instances

(b) Learning the Cardinality Model

(c) Computing Cardinality Kernels for all Bag Pairs

(d) Using a Kernel Method to Train a Classifier

Figure 2. The high-level scheme of the proposed kernel method for bag classification.

A graphical representation of the model is shown in Fig. 2(b). In our framework, we call this the *"Cardinality Model"*, and the details of its components are described as follows:

**Cardinality clique potential $\phi^C(Y, \mathbf{y})$:** a clique potential over all the instance labels and the bag label. This is used to model multi-instance or label proportion assumptions and is formulated as $\phi^C(Y, \mathbf{y}) = C^{(Y)}(\sum_i y_i)$. $C^{(+1)}$ and $C^{(-1)}$ are cardinality potentials for positive and negative bag labels, and in general could be expressed by any cardinality function. In this paper we work with the "normal" model in (3) and the "ratio-constrained" model in (4).

$$\begin{aligned} C^{(+1)}(c) &= \exp\left(-(\frac{c}{m} - \mu)^2/2\sigma^2\right) \\ C^{(-1)}(c) &= \exp\left(-(\frac{c}{m})^2/2\sigma^2\right). \end{aligned} \tag{3}$$

$$\begin{aligned} C^{(+1)}(c) &= \mathbb{1}(\frac{c}{m} >= \rho) \\ C^{(-1)}(c) &= \mathbb{1}(\frac{c}{m} < \rho). \end{aligned} \tag{4}$$

The parameter $\mu$ in the normal model or $\rho$ in the ratio-constrained model controls the proportion of positive labeled instances in a bag. The Normal model does not impose hard constraints on the number of positive instances, and consequently a positive bag can have any proportion of positive instances but it is more likely to be around $\mu$. On the other hand, the ratio-constrained model makes a hard constraint, assuming a bag must have at least a certain ratio ($\rho$) of positive instances.

**Instance-label potential $\phi^I_{\boldsymbol{\theta}}(\mathbf{x}_i, y_i)$:** represents the potential between each instance and its label. Essentially, this potential describes how likely it is for an instance (e.g. video frame) to receive a certain label (e.g. relevant or not to an event). It is parameterized as:

$$\phi^I_{\boldsymbol{\theta}}(\mathbf{x}_i, y_i) = \exp\left(\boldsymbol{\theta}^t \mathbf{x}_i \, y_i\right) \tag{5}$$

With these potential functions, the joint probability in (2) can be rewritten as

$$P(Y, \mathbf{y}|\mathbf{X}; \boldsymbol{\theta}) \propto C^{(Y)}(\sum_i y_i) \prod_i \exp\left(\boldsymbol{\theta}^t \mathbf{x}_i \, y_i\right). \tag{6}$$

And finally, the bag label likelihood, is obtained by

$$P(Y|\mathbf{X}; \boldsymbol{\theta}) = \sum_{\mathbf{y}} P(Y, \mathbf{y}|\mathbf{X}; \boldsymbol{\theta}) = \frac{Z^{(Y)}}{\sum_{Y'} Z^{(Y')}}, \tag{7}$$

$$\text{where } Z^{(Y)} = \sum_{\mathbf{y}} \left( C^{(Y)}(\sum_i y_i) \prod_i \exp\left(\boldsymbol{\theta}^t \mathbf{x}_i \, y_i\right) \right) \tag{8}$$

is the partition function of a standard cardinality potential model, which can be computed efficiently.

In summary, we have a unified probabilistic model that states the probability that a bag (e.g. video) receives a label based on classifying individual instances (e.g. frames), and a cardinality potential that prefers certain counts of positively labeled instances.

#### 3.1.1 Parameter Learning

Since only the bag labels, and not the instance labels, are provided in training, this Cardinality Model is a hidden conditional random field (HCRF). A commonly used algorithm for parameter learning is maximum a posteriori estimation of the parameters given the parameter prior distributions by maximizing the log likelihood function:

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_i \log P(Y_i|\mathbf{X}_i; \boldsymbol{\theta}) - \lambda \, r(\boldsymbol{\theta}). \tag{9}$$

This is maximum likelihood optimization of a HCRF with parameter regularization ($r(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_n$ for $L_n$-norm regularization). Gradient ascent is used to find the optimal parameters, where the gradients are obtained efficiently in terms of marginal probabilities [24].

### 3.2. Cardinality Kernel

This section presents the proposed probabilistic kernel for multi-instance classification. Kernels operate over a pair of inputs, in this case two bags. This kernel is defined using the Cardinality Models defined above. Each bag has its own set of instances, and a probabilistic model is defined over

each bag. A kernel over bags is formed by marginalizing over latent instance labels.

Given two bags $\mathbf{X}_p$ and $\mathbf{X}_q$, a joint kernel is defined between the combined instance features and instance labels for these bags $\mathbf{z}_p = (\mathbf{X}_p, \mathbf{y}_p)$ and $\mathbf{z}_q = (\mathbf{X}_q, \mathbf{y}_q)$:

$$k_z(\mathbf{z}_p, \mathbf{z}_q) = \sum_{i=1}^{m_p} \sum_{j=1}^{m_q} k_x(\mathbf{x}_{pi}, \mathbf{x}_{qj}) k_y(y_{pi}, y_{qj}), \quad (10)$$

where $k_x(\cdot,\cdot)$ is a standard kernel between single instances, and $k_y(\cdot,\cdot)$ is a kernel defined on discrete instance labels[1]. By marginalizing the joint kernel w.r.t. the hidden instance labels and with independence assumed between the bags, a kernel is defined on the bags as:

$$\tilde{k}(\mathbf{X}_p, \mathbf{X}_q) = \sum_{\mathbf{y}_p, \mathbf{y}_q} P(\mathbf{y}_p|\mathbf{X}_p) P(\mathbf{y}_q|\mathbf{X}_q) k_z(\mathbf{z}_p, \mathbf{z}_q). \quad (11)$$

Combining the fully observed label instance kernel (10) with the probabilistic version (11), it can be shown that the marginalized joint kernel is reduced to

$$\sum_{i=1}^{m_p} \sum_{j=1}^{m_q} \sum_{\mathbf{y}_p, \mathbf{y}_q} \Big( k_x(\mathbf{x}_{pi}, \mathbf{x}_{qj}) k_y(y_{pi}, y_{qj}) P(y_{pi}|\mathbf{X}_p) P(y_{qj}|\mathbf{X}_q) \Big). \quad (12)$$

In our proposed framework, $P(y_{pi}|\mathbf{X}_p)$ and $P(y_{qj}|\mathbf{X}_q)$ are obtained by

$$P(y_i|\mathbf{X}) = \sum_{Y} P(y_i|Y, \mathbf{X})\, P(Y|\mathbf{X}), \quad (13)$$

where $P(y_i|Y, \mathbf{X})$ are the marginal probabilities of a standard cardinality potential model, which can be computed efficiently in $O(m \log^2 m)$ time. Also $P(Y|\mathbf{X})$ is the bag label likelihood introduced in (7).

In general, any kernel for discrete spaces can be used as $k_y$. The most commonly used discrete kernel is $k_y(y_{pi}, y_{qj}) = \mathbb{1}(y_{pi} = y_{qj})$. However, since throughout this paper we are dealing with binary instance labels and we are interested in performing recognition with the most salient and positively relevant instances of a bag, $k_y$ is assumed to be

$$k_y(y_{pi}, y_{qj}) = \mathbb{1}(y_{pi} = 1) \cdot \mathbb{1}(y_{qj} = 1). \quad (14)$$

Using this, the kernel in (12) is simplified as:

$$\tilde{k}(\mathbf{X}_p, \mathbf{X}_q) = \sum_{i=1}^{m_p} \sum_{j=1}^{m_q} k_x(\mathbf{x}_{pi}, \mathbf{x}_{qj}) P(y_{pi} = 1|\mathbf{X}_p) P(y_{qj} = 1|\mathbf{X}_q). \quad (15)$$

[1] If $k_y(\cdot,\cdot)$ is set to 1, the resulting kernel will be equivalent to MI-Kernel [7]. Also, note that since the joint kernel is obtained by summing and multiplying the base kernels, it is proved to be a kernel, has all kernel properties, and can be safely plugged into kernel methods.

It is interesting to note that this kernel in (15) can be rewritten as

$$\tilde{k}(\mathbf{X}_p, \mathbf{X}_q) = \Big( \sum_{i=1}^{m_p} P(y_{pi} = 1|\mathbf{X}_p) \Psi(\mathbf{x}_{pi}) \Big) \Big( \sum_{j=1}^{m_q} P(y_{qj} = 1|\mathbf{X}_q) \Psi(\mathbf{x}_{qj}) \Big), \quad (16)$$

where $\Psi(\mathbf{x})$ is the mapping function that maps the instances to the underlying feature space of the instance kernel $k_x$. This proves that the unnormalized cardinality kernel in the original feature space corresponds to weighted sum of the instances in the induced feature space of $k_x$, where the weights are the marginal probabilities inferred from the Cardinality Model in the original space. It can be also shown that in the more general case of $k_y(y_{pi}, y_{qj}) = \mathbb{1}(y_{pi} = y_{qj})$, the resulting cardinality kernel would correspond to weighted sum of all the instances which take the same instance label in the mapped feature space and concatenating them altogether.

Finally, to avoid bias towards the bags with large numbers of instances, the kernel is normalized as [7]:

$$k(\mathbf{X}_p, \mathbf{X}_q) = \frac{\tilde{k}(\mathbf{X}_p, \mathbf{X}_q)}{\sqrt{\tilde{k}(\mathbf{X}_p, \mathbf{X}_p)} \sqrt{\tilde{k}(\mathbf{X}_q, \mathbf{X}_q)}}. \quad (17)$$

We call the resulting kernel the *"Cardinality Kernel"*. By using this kernel in the standard kernel SVM, we propose a method for multi-instance classification with cardinality relations.

### 3.3. Algorithm Summary

The proposed algorithm is summarized as follows. First the parameters $\boldsymbol{\theta}$ of the Cardinality Model are learned (Sec. 3.1). These parameters control the classification of individual instances and the cardinality relations for bag classification. Next, the marginal probabilities of instance labels under this model are inferred and used in the kernel function in (15). Finally, the kernel is normalized and plugged into an SVM classifier[2].

A comprehensive analysis of the computational complexity of the proposed algorithm can be found in the supplementary material. In short, the kernel in (15) can be evaluated in $O(m_p m_q d + m_p \log^2 m_p + m_q \log^2 m_q)$ time, where the basic kernel $k_x$ takes $O(d)$ time to compute, and the number of instances in each bag are $m_p$ and $m_q$.

## 4. Experiments

We provide empirical results on three tasks: group activity recognition, video event detection, and video interestingness analysis.

[2] For the parameter setting guidelines, see the supplementary material.

### 4.1. Collective Activity Recognition

The Collective Activity Dataset [6] comprises 44 videos (about 2500 video frames) of *crossing*, *waiting*, *queuing*, *walking*, and *talking*. Our goal is to classify the collective activity in each frame. To this end, we model the scene as a bag of people represented by the *action context* feature descriptors[3] developed in [18]. We use our proposed algorithms with the ratio-constrained cardinality model in (4) with $\rho = 0.5$, to encode a majority cardinality relation. We follow the same experimental settings as used in [18], i.e., the same 1/3 of the video clips were selected for test and the rest for training. The one-versus-all technique was employed for multi-class classification. We applied $l_2$-norm regularization in likelihood maximization of the Cardinality Model and simply used linear kernels as the instance kernels in our method. The results of our Cardinality Kernel are shown in Table 1 and compared with the following methods[4]: (1) SVM on global bag-of-words, (2) Graph-structured latent SVM method in [18], (3) MI-Kernel [7], (4) Cardinality Model of Section 3.1 (our own baseline).

Table 1. Comparison of classification accuracies of different algorithms on collective activity dataset. Both multi-class accuracy (MCA) and mean per-class (MPC) accuracy are shown because of class size imbalance.

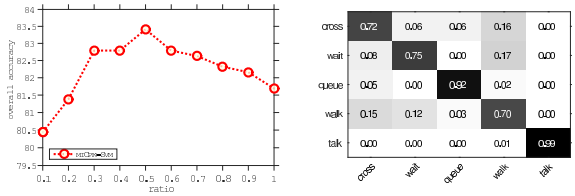
| Method | MCA | MPCA |
|---|---|---|
| Global bag-of-words with SVM [18] | 70.9 | 68.6 |
| Latent SVM with optimized graph [18] | 79.7 | 78.4 |
| Cardinality Model | 79.5 | 78.7 |
| MI-Kernel | 80.3 | 78.4 |
| Cardinality Kernel (our proposed method) | **83.4** | **81.9** |

Our simple Cardinality Model can achieve results comparable to the structure-optimized models by replacing spatial relations with cardinality relations. Further, the proposed Cardinality Kernel can significantly improve classification performance of the Cardinality Model. Finally, our Cardinality Kernel is considerably better than MI-Kernel, showing the advantage of using importance weights (i.e. probability of being positive) of each instance for non-uniform aggregation of instance kernels.

Fig. 3(a) illustrates the effect of $\rho$ in the ratio-constrained cardinality model on classification accuracy of the Cardinality Kernel. It can be seen that as expected, the best result is achieved with $\rho = 0.5$. We also provide the confusion matrix for the Cardinality Kernel method in Fig. 3(b). Finally, two examples of recognition with the Cardinality Model for crossing and waiting activities are visualized in Fig. 4.

[3]These features are based on a spatio-temporal context region around a person. So by using our cardinality-based model, the spatio-temporal and cardinality information are combined.

[4]All these methods follow the standard evaluation protocol introduced in [5]. See the supplementary material to find the comparison with the methods in [3, 2, 1], which use a different evaluation setting.




(a) (b)

Figure 3. Performance of the Cardinality Kernel on collective activity dataset. (a) Classification accuracy with different values of $\rho$ in the ratio-constrained cardinality model. (b) Confusion matrix with $\rho = 0.5$ (rows are the true labels, and columns are predicated labels)

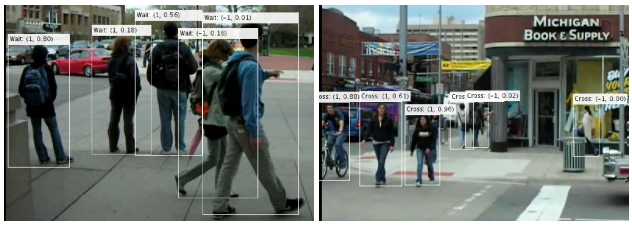



Figure 4. Examples of recognition with the proposed model. The annotation of each person shows the true activity label of the scene with a tuple, indicating the MAP-inferred action label and the corresponding marginal probability w.r.t. the the scene activity label. -1 values denote "not" of the corresponding category; people performing other actions (left: two people not waiting, right: people not crossing the street) are correctly given -1 labels.

### 4.2. Event Detection

We evaluate our proposed method for event detection on the TRECVID MED11 dataset [21]. Because of temporal clutter in the videos, not all parts of a video are relevant to the underlying event, and the video segments might have unequal contributions to event detection. Our framework can deal with this temporal ambiguity, i.e., when the evidence of an event is occurring in a video and what the degree of discrimination or importance of each temporal segment is. We represent each video as a bag of ten temporal video segments, where each segment is represented by pooling the features inside it. As the cardinality potential, we use the Normal model in (3) with $\mu = 1$ and $\sigma = 0.1$ to embed a soft and intuitive constraint on the number of positive instances: *the more relevant segments in a video, the higher the probability of occurring the event.*

We follow the evaluation protocol used in [29, 25]. The DEV-T split of MED11 dataset is used for validation and finding the hyper-parameters such as the regularization weights in learning the Cardinality Model and SVM. Then, we evaluate the methods on the DEV-O test collection (32061 videos), containing the events 6 to 15 and a large number of null (or background events). For training,

an Event-Kit collection of roughly 150 videos per event is used, and as in [29, 25], the classifiers are trained for each event versus all the others.

We compare our methods with the kernelized latent SVM methods in [29], applied to a structured model where the temporal location and scene type of the salient video segments are modeled as latent variables. To have a fair comparison, we use the same set of features: HOG3D, sparse SIFT, dense SIFT, HOG2x2, self-similarity descriptors (ssim), and color histograms, which are simply concatenated to a single feature vector[5]. For training the Cardinality Model regularized maximum likelihood is used with $l_1$-norm regularization, and for the Cardinality Kernel histogram intersection kernel is plugged as the instance kernel. The results in terms of average precision (AP) are shown in Fig. 5. It can be observed that based on mean AP, our proposed Cardinality Kernel clearly outperforms the baselines:

- The Cardinality Model of Sec. 3.1.
- Kernelized SVM (KSVM) and multiple kernel learning SVM (MKL-SVM), which are kernel methods with global bag-of-words models.
- MI-Kernel [7], which is a multi-instance kernel method with uniform aggregation of the instance kernels.

On the other hand, our method is comparable to the kernelized latent SVM (KLSVM) methods in [29]. However, our model is considerably less complicated, and unlike these methods, our proposed framework has exact and efficient inference and learning algorithms. For example the training time for our method is about 35 minutes per event, but those methods takes about 30 hours per event[6]. In addition, based on comparison on individual events, our proposed method achieves the best AP in 6 out of 10 events.

Recently, Lai et al. [17] proposed a multi-instance framework for video event detection, by treating a video as a bag of temporal video segments of different granularity. Since this is the closest work to ours, we run another experiment on TRECVID MED11 to evaluate performance of our algorithm compared to [17]. We use exactly the same settings as before, but since Lai et al. [17] used dense SIFT features, we also extract dense SIFT features quantized into a 1500-dimensional bag-of-words vector for each video segment[7], where the video segments are given by dividing each video into 10 equal parts. This is slightly different from the multi-granular approach in [17], where both the single frames and temporal video segments are used as the instances (single–g $\propto$SVM uses only single frames and multi–g $\propto$SVM uses both the single frames and video segments). The results are shown in Table (2). Our method outperforms multi–g $\propto$SVM (which is the best in [17]) by around 20%. In addition, our algorithm is more efficient, and training takes only about half an hour per event.

Table 2. Comparing our proposed Cardinality Kernel method with $\propto$SVM algorithms in [17] on TRECVID MED11. The best AP for each event is highlighted in bold

| Event | single–g $\propto$SVM [17] | multi–g $\propto$SVM [17] | Cardinality Kernel |
|---|---|---|---|
| 6 | 1.9 % | **3.8** % | 2.8 % |
| 7 | 2.6 % | **5.8** % | **5.8** % |
| 8 | 11.5 % | 11.7 % | **17.0** % |
| 9 | 4.9 % | 5.0 % | **8.8** % |
| 10 | 0.8 % | 0.9 % | **1.3** % |
| 11 | 1.8 % | 2.4 % | **3.4** % |
| 12 | 4.8 % | 5.0 % | **10.7** % |
| 13 | 1.7 % | 2.0 % | **4.7** % |
| 14 | 10.5 % | **11.0** % | 4.9 % |
| 15 | 2.5 % | **2.5** % | 1.4 % |
| mAP | 4.3 % | 5.0 % | **6.1** % |

### 4.3. Video Summarization by Detecting Interesting Video Segments

Recently, Gygli et al. [10] proposed a novel method for creating summaries from user videos by selecting a subset of video segments, which are interesting and informative. For this purpose, they created a benchmark dataset (SumMe[8]) of 25 raw user videos, summarized and annotated by 15 to 18 human subjects. In their proposed method, each video segment is scored by summing the *interestingness* score of its frames, estimated by a regression model learned from human annotations. At the end, a subset of video segments is selected such that the summary length is 15% of the input video.

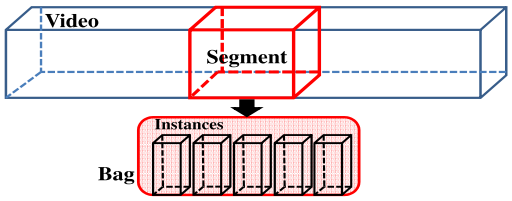



Figure 6. Detecting interesting video segments. A video is modeled as a bag of sub-segments.

In this paper, we propose a new approach for creating segment-level summaries. Instead of predicting the per-frame scores and using a heuristic aggregation operation such as "sum", we use our multi-instance model to directly

[5]In the experiments of this section we compare our method with the most relevant methods, which use the same features. By using, combining, or fusing other sets of features, better results can be achieved (e.g. [26, 31])

[6]We performed our experiments on an Intel(R) Core(TM) i7-2600 CPU @ 3.40GHz, and compared to our previous work [29].

[7]We use VLFeat, as in [17], though with fewer codewords (5000 in [17]). See the supplementary material for the results with more codewords.

[8]The dataset and evaluation code for computing the f-measure are available at `http://www.vision.ee.ethz.ch/~gyglim/vsum/`

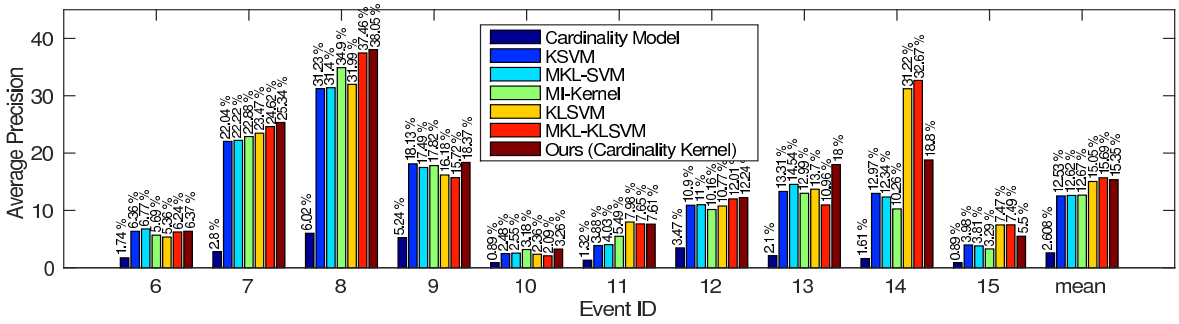



Figure 5. The APs for events 6 to 15 in TRECVID MED 2011. The results for KSVM, MKL-SVM, KLSVM, and MKL-KLSVM are reported from [29]. MI-Kernel is based on our own implementation of the algorithm in [7].

estimate the interestingness of a video segment. The proposed approach is illustrated in Fig. 6. Each segment is modeled as a bag of sub-segments, where a positive bag is a segment which has large overlap with human annotated summaries. To represent each sub-segment, we extract HSV color histogram (with $8 \times 8$ bins) and bag-of-words dense trajectory features [30] (with 4000 words) for each frame and max-pool the features over the sub-segment. Here, we summarize our method and the baselines:

- Ours: A segment is divided into 5 sub-segments, and the proposed Cardinality Kernel with Normal cardinality potential ($\mu = 1, \sigma = 0.1$) is used to score the segments.
- Global Model: A global representation of each segment is constructed by max-pooling the features inside it, and an SVM is trained on the segments.
- Single-Frame SVM: An SVM is trained on the frames, and the score of each segment is estimated by summing the frame scores.
- Single-Frame SVR: This is our simulation of the algorithm in [10] but with our own features, fixed length segments, and using support vector regression.

The top scoring $15\%$ of segments are selected in each.

For all methods a video is segmented into temporal segments of length $P_l = 1.85$ seconds (the segment length given in [10]), and histogram intersection kernel is used for training the SVMs. To evaluate the methods, the procedure in [10] is used: leave-one-out validation, comparison based on per segment f-measure. The results are shown in Fig. 7. It can be observed that our method outperforms the baselines and is competitive with the state-of-the-art results in [10]. In fact, although we are using general features (color histogram and dense trajectory) we achieve a performance which is comparable to the performance in [10], which uses specialized features to represent *attention*, *aesthetics*, *landmarks*, etc. Note that the best f-measure in [10] is obtained by over-segmenting a video into cuttable segments called *superframe*, using guidelines from editing theory.

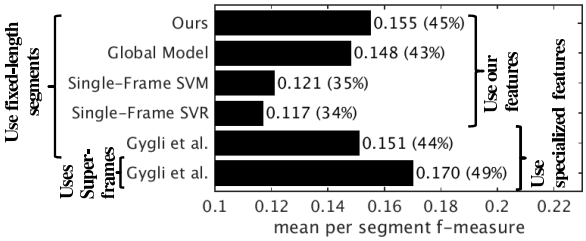



Figure 7. Comparison of different algorithms for segment-level summarization of the SumMe benchmark videos. The percent scores are relative to the average human.

## 5. Conclusion

We demonstrated the importance of cardinality relations in visual recognition. To this end, a probabilistic structured kernel method was introduced. This method is constructed based on a multi-instance cardinality model, which can explore different levels of ambiguity in instance labels and model different cardinality-based assumptions. We evaluated the performance of the proposed method on three challenging tasks: collective activity recognition, video event detection, and video summarization. The results showed that encoding cardinality relations and using a kernel approach with non-uniform (or probabilistic) aggregation of instances leads to significant improvement of classification performance. Further, the proposed method is powerful, straightforward to implement, with exact inference and learning, and can be simply integrated with off-the-shelf structured learning or kernel learning methods.

## References


[1] M. R. Amer, P. Lei, and S. Todorovic. Hirf: Hierarchical random field for collective activity recognition in videos. In *European Conference on Computer Vision (ECCV)*, pages 572–585. Springer, 2014. 6

[2] M. R. Amer, S. Todorovic, A. Fern, and S.-C. Zhu. Monte carlo tree search for scheduling activity recognition. In *International Conference on Computer Vision*, 2013. 6

[3] M. R. Amer, D. Xie, M. Zhao, S. Todorovic, and S. C. Zhu. Cost-sensitive top-down/bottom-up inference for multiscale activity recognition. In *European Conference on Computer Vision (ECCV)*, 2012. 1, 2, 6

[4] P. Bojanowski, R. Lajugie, F. Bach, I. Laptev, J. Ponce, C. Schmid, and J. Sivic. Weakly supervised action labeling in videos under ordering constraints. In *European Conference on Computer Vision (ECCV)*, 2014. 2

[5] W. Choi and S. Savarese. A unified framework for multi-target tracking and collective activity recognition. In *European Conference on Computer Vision (ECCV)*, pages 215–230. 2012. 1, 2, 6

[6] W. Choi, K. Shahid, and S. Savarese. What are they doing?: Collective activity classification using spatio-temporal relationship among people. In *9th International Workshop on Visual Surveillance*, 2009. 1, 2, 6

[7] T. Gärtner, P. Flach, A. Kowalczyk, and A. Smola. Multi-instance kernels. In *International Conference on Machine Learning (ICML)*, pages 179–186, 2002. 3, 5, 6, 7, 8

[8] A. Gupta, P. Srinivasan, J. Shi, and L. S. Davis. Understanding videos, constructing plots: Learning a visually grounded storyline model from annotated videos. In *Computer Vision and Pattern Recognition (CVPR)*, 2009. 1, 2

[9] R. Gupta, A. Diwan, and S. Sarawagi. Efficient inference with cardinality-based clique potentials. In *International Conference on Machine Learning (ICML)*, pages 329–336. ACM, 2007. 3

[10] M. Gygli, H. Grabner, H. Riemenschneider, and L. Van Gool. Creating summaries from user videos. In *European Conference on Computer Vision (ECCV)*, pages 505–520. Springer, 2014. 1, 2, 7, 8

[11] H. Hajimirsadeghi, J. Li, G. Mori, M. Zaki, and T. Sayed. Multiple instance learning by discriminative training of markov networks. In *Uncertainty in Artificial Intelligence (UAI)*, 2013. 3

[12] H. Hajimirsadeghi and G. Mori. Multiple instance real boosting with aggregation functions. In *International Conference on Pattern Recognition (ICPR)*, 2012. 3

[13] S. Khamis, V. I. Morariu, and L. S. Davis. Combining per-frame and per-track cues for multi-person action recognition. In *European Conference on Computer Vision (ECCV)*, 2012. 2

[14] A. Khosla, R. Hamid, C. Lin, and N. Sundaresan. Large-scale video summarization using web-image priors. In *Computer Vision and Pattern Recognition (CVPR)*, 2013. 2

[15] G. Kim, L. Sigal, and E. Xing. Joint summarization of large-scale collections of web images and videos for storyline reconstruction. In *Computer Vision and Pattern Recognition (CVPR)*, 2014. 2

[16] J. T. Kwok and P.-M. Cheung. Marginalized multi-instance kernels. In *International Joint Conferences on Artificial Intelligence (IJCAI)*, pages 901–906, 2007. 3

[17] K.-T. Lai, F. X. Yu, M.-S. Chen, and S.-F. Chang. Video event detection by inferring temporal instance labels. In *Computer Vision and Pattern Recognition (CVPR)*, 2014. 2, 3, 7

[18] T. Lan, Y. Wang, W. Yang, S. N. Robinovitch, and G. Mori. Discriminative latent models for recognizing contextual group activities. *IEEE Trans. Pattern Analysis and Machine Intelligence (T-PAMI)*, 34(8):1549–1562, 2012. 1, 2, 6

[19] W. Li, L. Duan, D. Xu, and I. Tsang. Text-based image retrieval using progressive multi-instance learning. In *International Conference on Computer Vision (ICCV)*, 2011. 3

[20] J. C. Niebles, C.-W. Chen, and L. Fei-Fei. Modeling temporal structure of decomposable motion segments for activity classification. In *European Conference on Computer Vision (ECCV)*, 2010. 2

[21] P. Over, G. Awad, J. Fiscus, B. Antonishek, M. Michel, A. F. Smeaton, W. Kraaij, G. Quénot, et al. Trecvid 2011-an overview of the goals, tasks, data, evaluation mechanisms and metrics. In *TRECVID 2011-TREC Video Retrieval Evaluation Online*, 2011. 1, 2, 6

[22] H. Pirsiavash and D. Ramanan. Parsing videos of actions with segmental grammars. In *Computer Vision and Pattern Recognition (CVPR)*, 2014. 2

[23] D. Potapov, M. Douze, Z. Harchaoui, and C. Schmid. Category-specific video summarization. In *European Conference on Computer Vision (ECCV)*, 2014. 2

[24] A. Quattoni, S. Wang, L.-P. Morency, M. Collins, and T. Darrell. Hidden conditional random fields. *IEEE Trans. on Pattern Analysis and Machine Intelligence (T-PAMI)*, 29(10):1848–1852, 2007. 4

[25] K. Tang, L. Fei-Fei, and D. Koller. Learning latent temporal structure for complex event detection. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1250–1257, 2012. 1, 2, 6, 7

[26] K. Tang, B. Yao, L. Fei-Fei, and D. Koller. Combining the right features for complex event recognition. In *International Conference on Computer Vision (ICCV)*, pages 2696–2703. IEEE, 2013. 2, 7

[27] D. Tarlow, K. Swersky, R. Zemel, R. Adams, and B. Frey. Fast exact inference for recursive cardinality models. In *Uncertainty in Artificial Intelligence (UAI)*, 2012. 3

[28] Y. Tian, R. Sukthankar, and M. Shah. Spatiotemporal deformable part models for action detection. In *Computer Vision and Pattern Recognition (CVPR)*, 2013. 1, 2

[29] A. Vahdat, K. Cannons, G. Mori, S. Oh, and I. Kim. Compositional models for video event detection: A multiple kernel learning latent variable approach. In *International Conference on Computer Vision (ICCV)*, pages 1185–1192, 2013. 1, 2, 6, 7, 8

[30] H. Wang, A. Kläser, C.Schmid, and C.-L. Liu. Action recognition by dense trajectories. In *Computer Vision and Pattern Recognition (CVPR)*, 2011. 8

[31] Z. Xu, I. W. Tsang, Y. Yang, Z. Ma, and A. G. Hauptmann. Event detection using multi-level relevance labels and mul-

tiple features. In *Computer Vision and Pattern Recognition (CVPR)*, pages 97–104. IEEE, 2014. 2, 7

[32] F. X. Yu, D. Liu, S. Kumar, T. Jebara, and S.-F. Chang. $\propto$ svm for learning with label proportions. In *International Conference on Machine Learning (ICML)*, 2013. 3

[33] Y. Zhu, N. Nayak, and A. Roy-Chowdhury. Context-aware modeling and recognition of activities in video. In *Computer Vision and Pattern Recognition (CVPR)*, 2013. 2