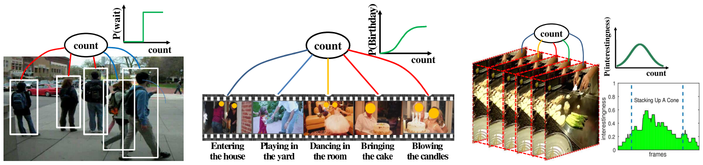
Visual Recognition by Counting Instances: A Multi-Instance Cardinality Potential Kernel

Hossein Hajimirsadeghi, Wang Yan, Arash Vahdat, Greg Mori
School of Computing Science, Simon Fraser University, Canada.

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 directed by counting the actions of individual people. Encoding these cardinality relationships can reduce sensitivity to clutter, in the form of irrelevant frames in a video or individuals not involved in group activity. This paper develops a powerful and flexible framework to embed 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.



(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 task [1]. 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 [3]. 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 [2]. The left image shows an important segment, where the chef is stacking up a cone. The right image shows the human-judged interestingness 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.

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.

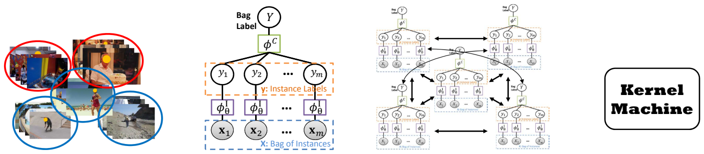
- **Showing the importance of cardinality relations for visual recognition.** We show in different applications that encoding cardinality relations, either hard (e.g. *majority*) or soft (e.g. *the more, the better*), can help to enhance recognition performance and increase robustness against labeling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

The proposed 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).

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.



(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.

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

[2] Michael Gygli, Helmut Grabner, Hayko Riemenschneider, and Luc Van Gool. Creating summaries from user videos. In *European Conference on Computer Vision (ECCV)*, pages 505–520. Springer, 2014.

[3] Paul Over, George Awad, Jon Fiscus, Brian Antonishek, Martial Michel, Alan F Smeaton, Wessel Kraaij, Georges 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.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.