A GENERALIZED MULTIPLE INSTANCE LEARNING ALGORITHM FOR ITERATIVE DISTILLATION AND CROSS-GRANULAR PROPAGATION OF VIDEO ANNOTATIONS

Feng Kang *

Michigan State University
East Lansing, MI

Milind R. Naphade, †

IBM Thomas J. Watson Research Center
Hawthorne, NY 10532

ABSTRACT

Video annotation is an expensive but necessary task for most vision and learning problems that require building models of visual semantics. This annotation gets prohibitively expensive especially when annotation has to happen at finer grained levels of regions in the videos. One way around the finer grained annotation dilemma is to support annotation at coarser granularity and then propagate this annotation to the finer granularity in a concept-dependent way. In this paper we propose a new generalized multiple instance learning algorithm that can work with any underlying density modeling techniques, and help propagate semantic concepts provided at the coarse granularity of video key-frames to finer grained regions. Our experiments on the NIST TRECVID common annotation corpus reveal improvement in annotation propagation accuracy between 3% to a dramatic 161%.

Index Terms— Algorithm, Information retrieval, Multiple Instance Learning, Video annotation.

1. INTRODUCTION

Semantic video indexing and search is a topic that is of great interest to the computer vision community and also a subject of evaluation benchmarks such as the NIST TRECVID benchmark [1]. As part of the TRECVID [1] benchmark, annotators across research organizations voluntarily annotated large broadcast news video corpora in 2003 and 2005. The idea was to create this common annotation to enable the modeling of a large number of semantic concepts (such as *Face*, *People*, *Sky*, *Road*, *Vehicle*, *Indoors*, *Outdoors*, etc.) that can then be used for enabling semantic visual search. In the 2003 common annotation exercise, researchers annotated both frame-level (*Outdoors, Indoors*, etc.) and regional concepts (*Sky, Face* etc.). For regional concepts, annotators also placed a rectangular bounding box around the region of interest. However in 2005, the common annotation task was only confined to providing frame-level annotations for both frame-level as well as region-level concepts. The main reason for this partial and incomplete annotation was the tremendous amount of time and effort it took to draw bounding boxes and annotate regional concepts. This large scale annotation and the decision to avoid placing bounding boxes underlines the key problem we attempt to solve in this paper.

Video annotation is an expensive task and region-level annotation makes it prohibitively expensive as seen from the largest voluntary annotation effort in TRECVID [1]. However the quality of detection based on regional ground truth is clearly superior to the quality that can be obtained by conventional learning techniques over keyframe-based ground truth. In the TRECVID 2003 annotation, annotators marked regions of interest corresponding to regional concepts with rectangular bounding boxes. Most the supervised learning algorithms can then be applied to this cleaner annotation of the concepts to the regions. This manual bounding box based annotation was time consuming and negatively impacted overall annotation quality. So while bounding boxes provided cleaner annotation where they were marked up, annotation fatigue took its toll on annotators who ended up missing a large number of concepts all together from keyframes that contained them and thus performed terribly on recall. This ended up hurting the performance of the TRECVID benchmark in 2003 and thus in 2005 the annotation was confined only to keyframe-level. This led to a significant improvement in annotation precision and recall.

In this paper we extend a successful generalized multiple instance learning algorithm [2] to get better region level ground truth through iterative distillation and cross-granular propagation of video annotations. This allows annotators to provide keyframe based annotations only at the global level and lets the algorithm deal with propagating it to the appropriate region within each keyframe. Using the TRECVID 2003 common annotation corpus, we conduct several experiments for a number of semantic concepts annotated. We then evaluate our approach using annotations that are available at both the keyframe level and the region level. Different from work in [2], the framework in this paper focuses more on the refining of ground truth and also includes smart selection and

*This work was performed by the author at the IBM Thomas J. Watson Research Center

†This material is based upon work funded in part by the U. S. Government. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the U.S. Government.

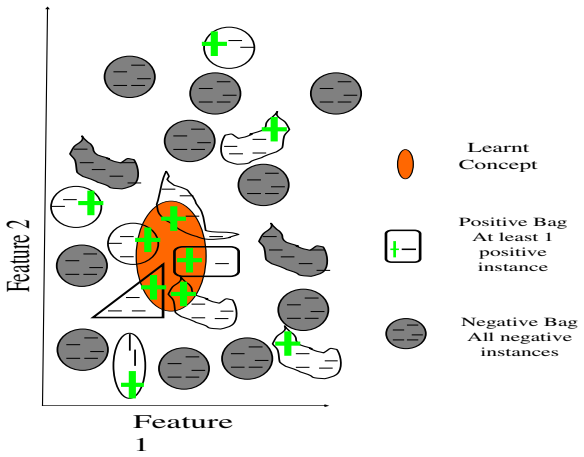

Fig. 1. A conceptual illustration of multiple instance learning at work. *Bags* are coarse level containers of instances that are targetted. The target feature space to be learnt is the blob in the middle close to as many positive bags and far from as many negative bags as possible.

iterative distillation as new strategy. It also uses different evaluation metric fit for the refining procedure.

2. RELATED WORK

Low quality training data substantially hurts model performance. Some general schemes that deal with improving training data quality and annotation accuracy include [3, 4, 5] etc. Smola et al [3] handle this situation by adding a regularization term to penalize certain data. Leo et al [4] average the prediction through bagging. Anelia et al [5], build positive and negative models based on features extracted from the whole image and then apply a selection procedure based on performance to prune the training data by thresholding. However all these approaches assume a single instance setting where there is no granular ambiguity and no need to resolve which region or regions in an image actually correspond to the annotations that have been provided at the image/frame level.

Work in multiple instance learning on the other hand focuses on identifying the finer level ground truth through the coarse level ground truth. In multiple instance learning, the labels are tagged to a bag of instances. The bags thus provide the coarse grain containers of the instances that are the fine grained objects to which these annotations actually correspond. Bags are marked positive if any instance is a positive instantiation of the concept of interest. Bags are marked negative if no instance is positive. Figure 1 illustrates this using bags and instances embedded in a fictional 2 dimensional feature space.

Multiple instance learning has been applied to image retrieval[6, 7]. We have previously shown a generalized multiple instance learning algorithm [2] to be more effective than diverse density [7]. In this paper, we will further extend the algorithm in [2] to be used for refining the region-level annotations.

3. A GENERALIZED MULTIPLE INSTANCE LEARNING ALGORITHM FOR DISTILLATION OF TRAINING DATA

As discussed from previous section that multiple instance learning(MIL) provides a natural modeling of the problem of propagating the image/frame level ground truth to regional level ground truth. We now describe the proposed algorithm for iterative distillation and cross-granularity propagation of annotations based on MIL. Our generalized algorithm can work with any underlying regression or density modeling technique. It is based on the following observations:

- All instances in negatively annotated bags are negative: This is definitive information that can be used unlike the positive bags where there is ambiguity.
- Smart selection of positive instances from positive bags: Since we know that there is at least one positive instance in the positive bag, we want to maximally use this information.
- Top K Selection: Not all potentially positive instances are created equal. After the potential positive instances are selected from each bag, we only choose some of them as most reliable instances to improve the precision of the intrinsic model
- Iterative distillation: Once an intrinsic model is built to represent the positive instances it can be iteratively refined until saturation.

The algorithm works as follows:

- Initialization: We start by building an initial negative hypothesis model from all instances in negatively annotated bags. We also proceed to build an initial positive hypothesis model by relaxing the ambiguity constraint for positive bags and using all instances in the positive bags to build the initial positive hypothesis model. A likelihood ratio test of the two hypotheses models is then used to rank each instance in a positively annotated bag.
- Selection of Likely Positive Instances: We know that not all instances from positively labeled bags are positive. Most of them are negative instances. There are choices to select the most likely positive instances from each bag. One of the method is to define a threshold for the likelihood ratio and treat the instances with confidence lower than the threshold as noisy data [5]. In our proposed method, we use the top K selection based on three steps:

1. Rank all instances from positive bags based on their likelihood ratio score.
2. Pick out instances from the step 1 in descending order of ranking till the point is reached where at least 1 instance from every positive bag is retained.
3. Top K Selection: From the above retained list of step 2 of positive instances choose the top $k\%$ of the instances.

Step 1 corresponds to our knowledge about multiple instance learning that at least one of the instances in the positive bag belonging to positive instances yet there might be several. Step 2 actually uses the minimal of maximum confidence of instances among all the positive bags to be threshold. In addition to ensuring that each bag is represented in the top K selection procedure, this step also allows the system to automatically determine this threshold. Step 3 is where only part of the list created in step 2 is selected for refining the positive hypothesis model. The intuition is to create a cleaner training set and thus create a highly precise ground truth to refine the positive hypothesis model. We experimentally verify this and will show later that the determination of value of $k\%$ is concept dependent and has strong correlation with the prior probability of the concept.

- Distillation/Refinement of the Positive Hypothesis Model: Having selected the most likely positive instances across bags, we then proceed to refine the model of the positive hypothesis. The negative hypothesis is unchanged from its estimate during the initialization step.

We then iterate the selection and refinement steps until convergence or a fixed number of iterations.

Once terminated the algorithm also provides as output the most likely positive instance(s) from each bag. As in [2], this algorithm allows to integrate different supervised or unsupervised method to build the positive and negative models as long as they can produce a ranking list of instances based on the confidence values.

Typically for most semantic concepts, the number of negative bags is much larger than that of the positive bags. Thus the computationally intense part is the estimation of the negative model and in comparison, the multiple iterations of refinement of the positive hypothesis model do not pose much of a computational burden.

One more thing to emphasize is that we only use the frame level annotation to build our models. This is frame level annotated concepts are propagated to the regions based on our algorithm. After this, we use the region level ground truth to evaluate the performance.

4. EXPERIMENTS

Our experiments are based on the NIST TRECVID 2003 corpus. 28054 annotated keyframes become the bags in our experiments. Each keyframe is segmented into 1-5 regions based on the manually created bounding boxes. Each of the keyframe is annotated with several concepts. To build the models in our algorithm, we will only use keyframe level ground truth for all the modeling. After we pick out the regions for a concept, we use the region level ground truth to evaluate the accuracy. For each of the regions of interest bounded by a rectangular bounding box, a 166 dimensional HSV color-correlogram feature vector is extracted to represent the region. We use the following five semantic concepts from the TRECVID common annotation corpus for our experiments: *Road, Sky, Face, Building, Person*

We use the improvement in accuracy of region-level annotation propagation as our metric. For a concept i at iteration j, let d_j^i represent the number of instances for which the predicted annotation matches the ground truth and let g^i denote the number of true positive instances in the ground truth. The propagation accuracy is the defined as follows:

$$acc_j^i = \frac{d_j^i}{g^i} \tag{1}$$

The accuracy gain for concept i in iteration j is defined as the relative improvement over the baseline which is iteration 1, and is characterized by usage of all instances in positive bags for building the positive hypothesis model and all instances in negative bags used for building the negative hypothesis model:

$$accuracy\_gain_j^i = \frac{acc_j^i - acc_1^i}{acc_1^i} \tag{2}$$

The average accuracy gain across all concepts at the j^{th} iteration over a group of concepts is defined as:

$$avg\_gain_j = \frac{1}{n} \sum_{i \in conceptlist} accuracy\_gain_j^i \tag{3}$$

where n is the number of concepts for testing.

Our results show that for all concepts except *Face*, there is significant improvement in accuracy gain and the average accuracy gain after the 4^{th} iteration is more than 60 % excluding *Face*. To see the whole picture more specifically, we plot the performance of one concept *Person* in Figures 2. The baseline is generated based on usage of all instances in positive bags for the positive hypothesis model and instances in the negative bags with certain down-sampling for the negative hypothesis model, which is equivalent to select 100% of the positive instances.

To further investigate the reason for the performance improvement, we list the best performance in terms of accuracy

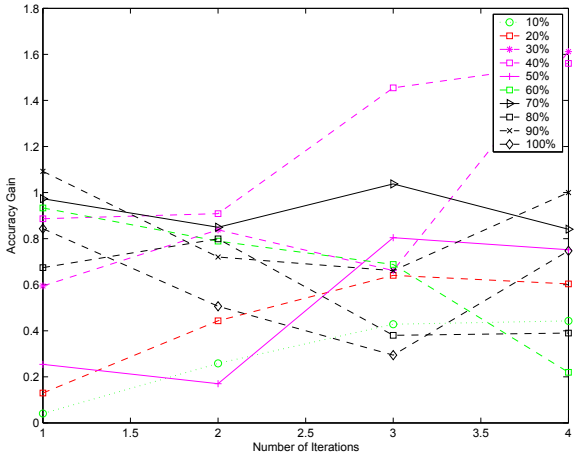

Fig. 2. Improvement of accuracy of cross-granular annotation propagation to regions for the concept *Person* for different top K selection percentages for 4 iterations of generalized multiple instance learning

	Road	Building	Sky	Person	Face
prior (%)	2.56	4.54	4.75	23.5	53.4
top K %	10	10	10	30	60
iterations	4	3	3	4	1
accuracy(%)	27.4	17.8	63.6	16.7	90.1
acc gain(%)	94	47.6	32.2	161	2.81

Table 1. Best Accuracy with Respect across number of iterations and top K selection percentages for different concepts

and accuracy gain across multiple iterations and multiple top K selection percentages juxtaposing it against the prior probability (expressed in terms of percentage of occurrence in the collection) of the concept in Table 1.

A concept like *Face* which occurs in more than 50 % of keyframes, and already has a very good propagation accuracy of 88% with even the baseline system cannot be expected to improve significantly beyond the 88% mark. Further with so many positive instances from so many positive bags, it is also difficult to expect that a small top K selection percentage will work for *Face* and that any new information will be added in this iterative distillation given the highly accurate starting point. As expected the optimal K for *Face* is 60 % and there is not much improvement beyond the first iteration.

On the contrary, for the rest of the concepts the exact opposite holds true. Three to four iterations are required to improve performance for the rest of the concepts. As for the top K selection percentage that seems to be strongly related to the prior probability of the concept. Thus for the relatively infrequent concepts *Road*, *Building Person*, and *Sky*, dramatic improvements in accuracy gain occur when the K is small and iteration is 3 or 4.

From the table, we see that, the proportion of data taken from the instances should increase as the prior probability increases. Another observation is that as the prior increases, the number of iterations required to achieve best performance usually decreases.

5. CONCLUSION

In this paper, we propose a generalized multiple instance learning framework to refine the training data. Our algorithm is based on three core concepts: likelihood ratio based ranking of candidate positive instances; smart selection of a limited number of candidate positive instances for refinement of positive hypothesis; and iterative refinement of the positive hypothesis model until convergence. We then apply the iterated positive hypothesis model and negative hypothesis model to propagate the keyframe level annotation to the most likely regional candidates from all the positively annotated keyframes. Using the TRECVID 2003 common annotation corpus that has region-level annotation ground truth, We show improvement in accuracy gain for all concepts tested with substantial gains for those concepts that start with low accuracy with the baseline system and exhibit infrequent probability of occurrence. We also verify empirically that the choice of best parameters is strongly related to prior probability of the concept.

6. REFERENCES

[1] "TREC Video Retrieval," 2003, National Institute of Standards and Technology, http://www-nlpir.nist.gov/projects/trecvid/.

[2] Milind Naphade and John Smith, "A generalized multiple instance learning algorithm for large scale modeling of multimedia semantics," in *2005 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2005.

[3] Bernhard Schlkopf and Alexander J. Smola, *Learning with kernels*, MIP Press, Cambridge, MA, 2002.

[4] Leo Breiman, "Bagging predictors," *Machine Learning*, 1996.

[5] Anelia Angelova, Yaser S. Abu-Mostafa, and Pietro Perona, "Pruning training sets for learning of object categories," in *CVPR*, 2005.

[6] Oded Maron and Aparna Lakshmi Ratan, "Multiple-instance learning for natural scene classification," in *ICML '98*, 1998.

[7] A. Ratan, O.Maron, W. Grimson, and T. Lozano-Perez, "A framework for learning query concepts in image classification," in *CVPR*, 1999.