In Defence of Negative Mining for Annotating Weakly Labelled Data

Parthipan Siva, Chris Russell*, and Tao Xiang

Queen Mary, University of London,
Mile End Road, London E1 4NS, United Kingdom
{psiva,chrisr,txiang}@eecs.qmul.ac.uk

Abstract. We propose a novel approach to annotating weakly labelled data. In contrast to many existing approaches that perform annotation by seeking clusters of self-similar exemplars (minimising intra-class variance), we perform image annotation by selecting exemplars that have never occurred before in the much larger, and strongly annotated, negative training set (maximising inter-class variance). Compared to existing methods, our approach is fast, robust, and obtains state of the art results on two challenging data-sets – VOC2007 (all poses), and the MSR2 action data-set, where we obtain a 10% increase. Moreover, this use of negative mining complements existing methods, that seek to minimize the intra-class variance, and can be readily integrated with many of them.

Keywords: Weakly Supervised Learning, Multiple-Instance Learning, Negative Mining, Automatic Annotation.

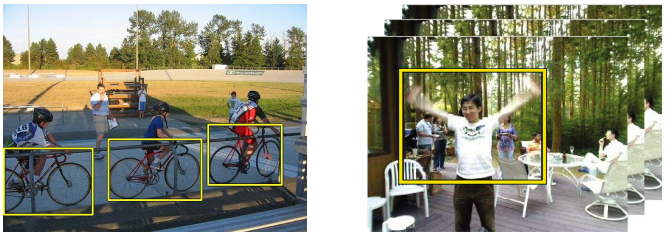
1 Introduction

Detecting objects in images [1, 2] and actions in videos [3, 4] are among the most widely studied computer vision problems, with applications in consumer photography, surveillance and automatic media tagging. Typically, these standard detectors are fully-supervised, that is they require a large body of training data where the location of the objects/actions in images/videos have been manually annotated, as shown in Fig.1. With the emergence of digital media, and the rise of high-speed internet, raw images and video are available for little to no cost. However, the manual annotation of object and action locations remains tedious, slow, and expensive, and as a result there has been a great interest in training detectors with weak supervision [5–8].

For weakly supervised training of object detector, each image in the training set is annotated with a weak label indicating if the image contains the object of interest or not, but not the locations of the object. As shown in Fig. 2, a weakly labelled data-set consists of two types of images: a set of weakly-labelled positive images where the exact location of object is unknown, and a set of strongly labelled negative images which we know for sure that every location in the image does not contain the object of interest.

* This author was funded by the European Research Council under the ERC Starting Grant agreement 204871-HUMANIS.

A. Fitzgibbon et al. (Eds.): ECCV 2012, Part III, LNCS 7574, pp. 594–608, 2012.



(a) Bicycle Annotation (b) Hand-waving Annotation

Fig. 1. Manual annotation of object and action data-sets, which are required for fully supervised detector training

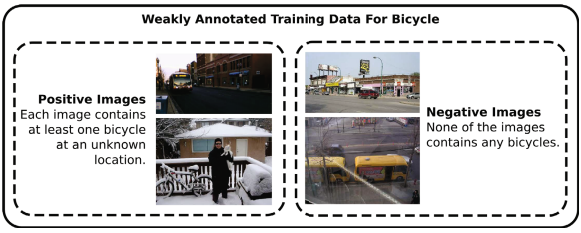


Fig. 2. In the annotating of weakly labelled data task we have a set of images or videos where the object or action of interest is present and a set of images or videos where the object or action is not present. Absence of object or action is strong information, as we know every part of the image or video is negative, whereas the presence of object or action is weak information as we do not know where the object or action is located.

Automatically annotating weakly-labelled training data is typically posed as a multiple-instance learning (MIL) problem [5, 6, 8]. Within a MIL framework, a single image weakly labelled with data such as: *"This image contains a bike."* is represented as a bag containing a set of instances, or possible locations of the object. *Positive bag* is used to refer an image containing at least one instance of the class, while *negative bags* are those that contain no positive instances. Given a set of positive and negative bags for training, the goal of MIL is to train a classifier that can correctly classify a test bag or test instance as either positive or negative. The latter is more relevant in the context of detector learning. In particular, taking a MIL approach, the problem of detector learning can be solved in two stages: in the first stage a decision is made as to which portion of the positive images represent the objects, and in the second stage a standard detector is trained from the decision made in the first stage. This approach has substantial pragmatic value, as it allows computationally intensive *transductive* [9] methods [8, 10] for automatically annotating the object location in the training data, and efficient state-of-the-art detectors [2] to locate the objects in the test data.

Classical MIL approaches [11, 12] make use of two different types of information to train a classifier: intra-class and inter-class. *Intra-class* information

concerns the selected positive instances. The information is typically exploited by enforcing that the selected positive instances look similar to each other. In contrast *inter-class* information refers to the difference in appearance between selected positive and negative instances. This information is normally used by introducing a constraint that all instances selected as positive look dissimilar to the instances selected as negative. In the case of automatic annotation of object locations [6, 8] a third type of information, *saliency*, is often exploited. *Saliency*[1] refers to knowledge about the appearance of foreground objects, i.e. things [13], regardless of the class of object to which they belong. Saliency may capture generic knowledge regarding the typical size and location of objects in photos, or express a relationship between the strength of image edges and the location of object bounding boxes [14]. Saliency is typically used to prune the space of possible object or action locations a priori, allowing us to consider a reduced set of possible locations. The measure of saliency itself can also be used for selecting positive instances.

In this paper, we ask a question: "Which of intra- and inter-class information is more useful in practice?". A close examination of a widely used benchmarking data-set can give us some hints. The VOC2007 data-set [15] is often used to evaluate object detectors and MIL approaches. A typical class in this data-set has approximately 300 associated images containing an object and 4,700 images not containing that object. If 100 candidate object locations (instances) per image are extracted using a saliency measure, this gives 470,000 strongly labelled negative instances vs. 300 true objects located somewhere within the 30,000 instances in the positive images. Furthermore, object locations proposed by the saliency measure may not include the true object location[2]. Therefore, when considering intra-class distances, we have $< \mathbf{300}$ unlabelled similar positive instances in a high-dimensional feature space vs. an inter-class distance based upon $\mathbf{470,000}$ strongly labelled instances in the same high-dimensional space. For object detection, the feature spaces tend to have thousands, or even hundreds of thousands of dimensions. When using an RBF kernel, or nearest neighbour classifier in such high dimensional feature spaces, the coverage provided by $470,000$ labelled negative instances provides substantially more useful inter information than the intra information provided by the 300 positive instances.

This simple observation motivates our approach, which only makes use of the inter-class information (strongly labelled negative examples), and is referred to as *negative mining*. By combining negative mining with existing saliency measures [14] we are able to produce a classifier that outperforms the majority of existing approaches to image and video annotation, and can be readily integrated with many of them.

In this paper we show 1) how strongly labelled negative data can be used to create a state-of-the-art classifier, bypassing the problem of resolving complex interdependencies between the positive bags and 2) how our classifier can be

[1] Typically learned from meta-data.

[2] On the VOC2007data-set, in 30% of the images the saliency measure of [14] fails to propose a valid location.

fused with several existing approaches to MIL-based image annotation [8, 16, 17], where it invariably leads to an improvement over the method it was fused with.

2 Prior Works

Traditional approaches to MIL [11, 12] were applied to image-level categorisation. We are interested in a related but different problem of selecting which instances in the positive training set are true positives, and our literature review focuses on those approaches which have been successfully applied to the annotation of weakly labelled data.

The MI-SVM formulation of Andrews et al. [17] makes use of both inter and intra-class information, as it seeks a linear classifier that maximise the margin between the positive and negative instances. Both [5] and [7] make use of this approach. They requires a set of initial positive instance and for that both [5] and [7] use the entire positive image as the initial positive instance. We make no such assumptions and will show that negative mining in conjuncture with a saliency measure can outperform these MI-SVM formulations. We also show how fusing negative mining and saliency measures with MI-SVM formulation can further improve annotation accuracy.

Another method to use only inter-class and intra-class information is by Siva and Xiang [16] for annotating weakly labelled action data. They also made use of the distance to the nearest negative example, however they combine this with intra-class measure into a single cost function. We show that our negative mining measure alone outperforms their combined cost function on all data-sets. Furthermore we propose the use of saliency which significantly improve the results on the MSR2 action data-set used in [16].

Deselaers et al. [6] and Siva and Xiang [8] made use of all three types of information: inter-class, intra-class and saliency. Deselaers et al. [6] only use their intra-class measure in the first iteration then use the selected instances to iteratively tune their choice of inter-class and saliency measures on an additional, manually-annotated, auxiliary data-set. Our experimental results strongly suggest that inter-class information is more reliable and would provide more useful information at initialisation. Siva and Xiang [8] treat inter-class and intra-class information independently from each other, then fuse the results at a score level. For their inter-class measure, they use the MI-SVM formulation like [5] and [7]. They then merge intra-class and saliency in a single cost function used to select a set of similar looking instances. Again we show that inter-class and saliency are more meaningful measures and that combining inter-class and saliency leads to greater accuracy than combining intra-class and saliency measures. We will also show that the results of [8] can be further improved by fusing with our approach.

Fu et al. [18] also made use of the distance of instances from the negative bags. They selected instances furthest from the negative bags and used them to initialise cluster centres, which were then used to create the bag level feature descriptors of [12]. Their goal was a bag-level classifier and we differ from them in that we are interested in the direct annotation of the instances in the training

set. Furthermore, unlike [18], we treat and evaluate negative mining as a classifier in its own right rather than as a pre-processing heuristic.

3 Methodology

A formal definition of negative mining is given in section 3.1. Section 3.2 contains a discussion of feature normalisation, which is essential for getting negative mining to work as a MIL method. Finally, we describe the saliency measure used to define instances in images and videos in section 3.3.

3.1 Negative Mining

Consider a data-set consisting of a set of positive images I_i^+ that contain the object of interest and a set of negative images I_i^- which do not. Following [6] and [8] we consider a set of 100 salient locations or instances $x_{i,j=1...100}$ in each image i. Each instance $x_{i,j}$ is represented by a bag-of-words (BOW) histogram. The goal is to select a single instance x_i^+ from each positive image I^+ corresponding to the correct location of an object of interest. The negative mining algorithm accomplishes this by selecting the instance that maximises the distance to the nearest neighbour in any image containing only negative instances $x_{i,j}^-$,

$$x_i^+ = \arg\max_{x_{i,j}^+} ||x_{i,j}^+ - N(x_{i,j}^+)||_1, \tag{1}$$

where $||\cdot||_1$ is the L_1 norm and $N(x_{i,j}^+)$ refers to the negative nearest neighbour of $x_{i,j}^+$. As mentioned earlier, for a typical VOC class, we have 470,000 negative instances ($x_{i=1...4700,j=1...100}^-$) in a $2,000$ dimensional space. As such, the key to an efficient algorithm is fast nearest neighbour look-up, and to handle the large volume of data efficiently we make use of a KD-tree based approximate nearest neighbour algorithm [19], with 16 trees.

This approach of mining the nearest negative instance relies on the abundance of known negative instances and unlike [6, 8] requires no optimisation of intra-class cost function, resulting in a computationally efficient algorithm. Furthermore, unlike [8], this inter-class measure does not assume that the average instance in each positive bag represents a positive instance, nor does it assume that the entire image is a positive instance like [5, 7].

An important variation on (1) incorporates a saliency measure. If we have a measure $\Phi(\cdot)$, which serves as a prior of how likely an instance is to be an object of interest, regardless of the choice of class, we can simply add $\Phi(x_{i,j}^+)$ to our negative mining cost:

$$x_i^+ = \arg\max_{x_{i,j}^+} \left(||x_{i,j}^+ - N(x_{i,j}^+)||_1 + \Phi(x_{i,j}^+)\right) \tag{2}$$

Later in section 3.3, we define $\Phi(x_{i,j}^+)$ for each of the data-sets considered.

3.2 Normalisation Strategies

Finding the instance with the maximum negative nearest neighbour (NNN) distance in a positive bag is complicated by the drastic change in the size of proposed instances. Consider Fig. 3(a) where we plot the distance from all the instances in the positive bags of a single class to its nearest negative neighbour (Distance $= ||x^+_{i,j} - N(x^+_{i,j})||_1$) vs. the instance's size in pixel ($sizeof(x^+_{ij})$). The plot uses standard normalised BOW histograms, which are typically used for both object and action detection:

$$\hat{h}(i) = \frac{h(i)}{\sum_{i=1}^{B} h(i)}, \tag{3}$$

where h is the BOW histogram composed of B bins, and the L_1 distance between them[3]. We observe that the NNN distance for instances small in size are almost always much greater than the NNN distance associated with instances large in size. This behaviour can be attributed to sampling artefacts: small boxes naturally contain fewer densely sampled words. Compared to a large box, the distribution of words associated with a small box is much more likely to have a few sharply peaked modes, and many empty bins.

As a consequence, when selecting the instances that maximise the normalised distance to the nearest negative instance we select very small instances in each positive bag. The same behaviour can be observed for other type of distances based upon normalised BOW histograms (Fig. 3(d)). In the VOC data-set, this bias is extremely inappropriate, as the majority of images contain large object instances, and on the VOC2007 data-set this causes negative mining to perform *ten times worse than the random selection of positive instances.* A similar degradation in performance can be observed on the MSR2 data-set [20].

On the contrary, if we use unnormalised histograms (Fig. 3(c)), together with the L_1 distance, we observe the opposite effect. Large instances contain many dense words, and owing to the sheer number of words, typically have a large distance from their nearest negative neighbour, while small instances lie very close to their NNN. Although biased towards large instances this measure is at least biased in the correct direction, and performs substantially better than random.

To minimise the bias towards either large or small boxes, we consider the novel measure of root-normalised histograms

$$\hat{h}(i) = \frac{h(i)}{\sqrt{\sum_{i=1}^{B} h(i)}}. \tag{4}$$

Empirically this measure performs better than either normalised or unnormalised histograms. We compare the accuracy of weak annotation using negative mining with the different nearest neighbour in section 4. Figure 3(b) shows the relationship of distance vs. instance size of root-normalised histograms.

[3] This is equivalent to the histogram distance $d(x, y) = 1 - \sum_k \min(x_k, y_k)$ proposed in [16].

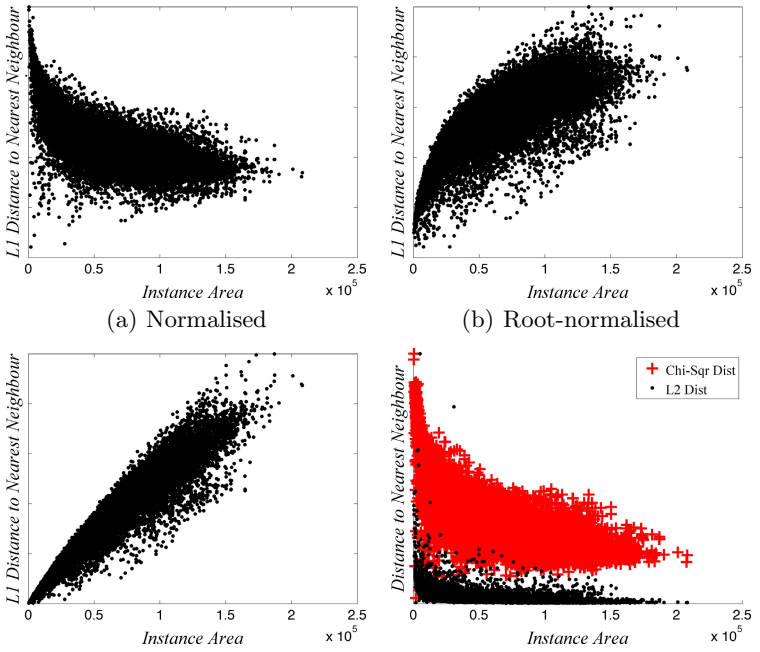


Fig. 3. The L_1 distances from all instances in the positive bags of the *cat* class in VOC2007 to the nearest negative instance plotted against the size in pixels of each instance in the positive bag. The L_1 distance is computed on a 2,000 word BOW histogram formed from regular grid SIFT features. *(a-c)* illustrate the effect of different normalisation strategies on BOW histograms. We see that selecting the instance with the largest distance to the nearest negative is correlated to the size of the instance and this correlation is less pronounced for root-norm. *(d)* shows that this phenomenon is not limited to the L_1 distance.

3.3 Saliency

Saliency is used at two places in our framework. We require it to propose a small set of viable instances or potential locations of objects and actions in each image or video. We also require a saliency measure of how likely a location is to be an object or action of any class, which we use in (2).

Instance Definition. To propose potential object locations on the VOC2007 data-set, we use the generic object detector [14] which was also used in other works on weak object class annotation [6, 8]. The first 100 samples from the generic object detector per image are used as instances. For direct comparison with [6, 8] we use version 1.01 of [14].

To propose potential locations of actions in the MSR2 video data-set we follow the procedure in [16]. We run a person detector [2] on each video frame and select a cuboid surrounding the detected person as a potential action location. From all these possible cuboids, 200 instances are selected based upon the spatio-temporal interest point (STIP) density.

Instance Measure. To measure how likely an instance $x_{i,j}$ is to be a positive location of any object ($\Phi(x^+_{i,j})$), we use the value of objectness returned by the generic object detector [14]. Objectness has also been used by [6] and [8] for this purpose.

For the action data-set, current literature does not have an equivalent of objectness. Instead we propose a simple heuristic, and choose $\Phi(x^+_{i,j}) = 0.6D$ where D is the density of STIPs in the defined cuboid. There are two reasons for using STIP density: first STIP density is used to sample the potential action cuboids (instances) from each video per [16] and, second, we expect motion where ever there is an action being performed which will generate a lot of STIPs. We weight STIP density by 0.6 as it is simple heuristic in comparison with objectness used for object detection.

4 Experiments

The data-sets and features used in our experiments are outlined in this section. Our main analysis is the comparison of our negative mining result with state-of-the-art results in section 4.1. Improvements in other MIL formulations when fused with the negative mining are given in section 4.2. Different histogram normalisation strategies are evaluated in section 4.3. Finally in section 4.4 we evaluate a leave-one-out classifier, assuming we have complete manual annotation of the training set, to determine the theoretical upper bound on the intra-class information in comparison to the inter-class information.

Data-Sets. For object detection we use the challenging Pascal VOC2007 data-set [15] and for action we use the MSR2 data-set [21]. For comparison with [8] we use VOCAll which consists of all 20 classes with *no* pose annotation, unlike [6] and [7] who use manual pose annotation. For comparison with [6] and [7] we also include the VOC6x2 which consists of six classes (aeroplane, bicycle, boat, bus, horse, and motorbike) where the left and right pose are considered as separate classes for a total of 12 classes. For both VOCAll and VOC6x2 we only exclude images annotated as difficult in the VOC2007 data-set. As in the other works [6], all results are presented as the percent of correct localisation, where correct localisation refers to 50% overlap of selected instance with ground truth as defined in [15].

Features. For features we use BOW histograms. For the VOC data-set BOW histograms are formed from regular grid SIFT features constructed with VLFEAT [22]. For the MSR2 data-set BOW histograms are formed using the histogram of flow

and histogram of gradient feature at spatio-temporal interest points (STIPs) defined in [23]. For both data-sets kmeans clustering is used to construct a 2,000 word codebook.

4.1 Comparison to State-of-the-Art

We report our comparison of negative mining with existing state-of-the-art image annotation algorithms in Table 1 (refer to Fig. 8 for more detailed results and Fig. 7 for illustrations). For the VOCAll the negative mining (N) alone (27.1%) outperforms both the inter-class (26.6%) and intra-class (23.9%) classifiers of Siva and Xiang [8]. When negative mining is combined with saliency ($N + \Phi$) our performance (29.0%) is slightly better than the combined inter- and intra-class method of [8] (28.9%).

For the restricted VOC6x2 data-set our negative mining (N) method (35.0%) is similar to saliency (Φ) alone (34.8%), and when combined ($N + \Phi$) the performance increases (37.1%). The combined method is better than [7] without their cropping heuristic and also better than [6] using a single feature. The inter-class measure proposed by [8] works better on the single pose subset than our negative mining based inter-class measure. However, as we show in section 4.2, the inter-class measure of [8] can be further improved by fusing its score with our inter-class measure. Through the use of post-processing [7], multiple features [6] and iteratively training object models [6–8] the annotation results can be further improved to 49%, 50%, and 61% for [6–8] respectively. We do not compare directly with these further refinements as our analysis is on the use of negative mining as a classifier for the initial annotation of the weakly annotated data. However, four rounds of iterative training of a detector using negative mining as initialisation results in an annotation accuracy of 46%.

More interestingly we note that the intra-class measure reported in [8] is actually intra-class with saliency as they included objectness value in their intra-class measure (Eq. 5 of [8]). However, the performance of their intra-class measure (23.9%) is worse than simply using objectness alone (25.1%) for the VOCAll set (Fig. 4) and the same (34.8%) for the VOC6x2 data-set. This clearly shows that their use of intra-class measure, at best does not improve the performance, and at worse significantly reduces the performance of saliency. In contrast, our inter-class measure consistently boosts the performance of saliency on all data-sets.

On the MSR2 data-set negative mining outperforms all existing methods. Our proposed saliency measure (STIP density) also outperforms MI-SVM [17] by itself but is not as strong as the methods of [16, 24]. However, saliency combined with negative mining gives a 10% performance boost over all pre-existing methods, including [24] which makes uses of a single manual annotation.

4.2 Fusion with Existing Classifiers

We combine negative mining and saliency with the intra-class, inter-class and combined inter-intra class measure of [8] as well as the intra-class measure of [16]. Each measure provides a score for individual instance in the positive bag and

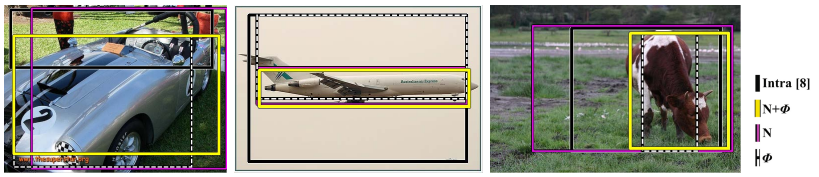


Fig. 4. Results using negative mining (N), saliency (Φ), and combined negative mining and saliency ($N + \Phi$) methods

Table 1. Results using negative mining (N), saliency (Φ), and combined negative mining and saliency ($N + \Phi$) methods

Data-set	Pandey [7]** Init. Only		Localisation Only [6]**		Siva and Xiang [8]*			MI-SVM [17]*	Nguyen et al. [5]*	Φ	N	N+Φ
	no crop	with crop	1 Feat.	5 Feat.	Intra $\mathcal{C}^*$	Inter $\mathcal{C}^\dagger$	Intra + Inter $\mathcal{C}'$					
VOCAll Avg.	N/A	N/A	N/A	N/A	23.9	26.6	28.9	25.4	22.4	25.1	27.1	29.0
VOC6x2 Avg.	36.7	43.7	35.0	39.0	34.8	39.0	39.6	38.7	24.9	34.8	35.0	37.1

*As reported in [8]. **As reported in [7] and [6] respectively.

Data-set	[16]	[24]	MI-SVM	Φ	N	N+Φ
MSR2 Avg.	71.2	71.7	55.8	60.8	73.9	81.5

Table 2. Augmention of existing methods by fusing with our negative mining (N) and saliency (Φ)

Data-set	Intra [8]	Intra [8] + N+Φ	Nguyen et al. [5] *MI-SVM Init. Img.*	Inter [8] *MI-SVM Init. Avg.*	Inter [8] + N+Φ	Inter [8]+ Intra [8]	Inter [8]+ Intra [8]+ N+Φ	Intra [16]	Intra [16]+ N+Φ
	One Iteration								
VOCAll Avg.	23.9	**28.8**	21.3	26.6	**29.6**	28.9	**30.2**	21.4	**24.3**
VOC6x2 Avg.	34.8	**36.5**	24.6	39.0	**40.8**	39.6	**40.2**	31.0	**36.2**

the instance that maximises the sum of all measure scores is selected as the true positive instance. We weight all measures equally. Average results are shown in Table 2 and more detailed per class results are in Table 4. Note that inter-class measure of [8] is MI-SVM [17] run for a single iteration and initialised with the average instance in each positive bag. Initialising with average instance performs better than initialising with the entire image, the method used by [5]. In all cases, we show that a score level fusion with our method (combined negative mining and saliency, $N + \Phi$) consistently boosts the performance of all other methods by between 1 and 5%.

4.3 Effect of Normalisation Strategies

As mentioned in section 3.2 the normalisation of the BOW histogram is vital for nearest neighbour to work. Fig. 5 shows annotation accuracy with normalised,

Data-set	VOCAll	VOC6x2	MSR2
Norm	1.4	2.3	14.7
Root-Norm	**27.1**	**35.0**	73.9
UnNorm	26.3	33.1	**74.3**
Rand	14.7	18.0	41.7

Fig. 5. Comparison of different normalisation strategies. See supplementary materials for per class results.

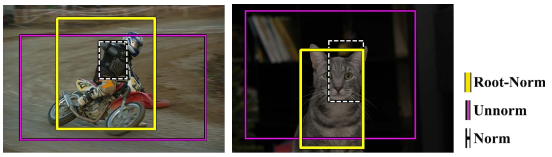


Fig. 6. The localisation result using just negative mining for the different BOW normalisation strategies. Normalised histogram tends to select small boxes, unnormalised tends to select large boxes, and root-normalised exhibits a weaker bias for large and small boxes.

Table 3. Leave-one-out results evaluating the usefulness of intra-class distance (P) when manually annotating all but one bag in comparison to negative mining (N) and saliency (Φ). See supplementary materials for per class results.

Data-set	Φ	N	$N+\Phi$	$-P$	$-P+N$	$-P+\Phi$	$-P+N+\Phi$
VOCAll Avg.	25.1	27.1	29.0	26.3	27.7	28.6	**29.6**
VOC6x2 Avg.	34.8	35.0	37.1	38.1	37.3	38.4	**39.2**
MSR2 Avg.	60.8	73.9	81.5	79.5	77.3	73.4	**82.6**

root-normalised and unnormalised BOW histograms and L_1 distance. We also show the performance of randomly selecting an instance from each positive bag.

The performance of normalised histogram for selecting the instance that maximises the distance to the nearest negative is substantially worse than random selection. This is because the use of normalised histograms actively selects the smallest instance (bounding box) in each positive bag which is not the behaviour you want for most bags (Fig. 6). Unnormalised histograms have the opposite effect, that is they tend to select the largest instance (bounding box). Root-normalised histograms tend to be a good compromise between normalised and unnormalised for all data-sets. Only exception is on the MSR2 data-set where the unnormalised distance is slightly better due to the fact there are no overly large instances proposed for the action data-set in comparison to the object data-set.

4.4 Leave-One-Out Positive Mining vs. Negative Mining

Our approach is based on the assumption that intra-class measure is not suitable for the first step in annotating weakly labelled data due to the availability of very few true positive data in a high dimensional space. To validate this we take a leave-one-out approach in which we assume all but one bag is manually annotated; this gives a theoretical upper bound on the intra-class information. In this scenario the unannotated bag can be annotated by selecting the instance that minimises the distance to its nearest manually annotated ground truth data; in practice we are maximising the negative distance to be consistent with (1). We also consider combining this leave-one-out intra-class distance with our negative mining and saliency measures. We use the normalisation strategies (see section 3.2) that maximise classification accuracy, being always *normalised* for

Fig. 7. Sucesses and failures of negative mining with saliency ($N + \Phi$)

Class	Siva and Xiang [8]* Intra $\mathcal{C}^*$	Siva and Xiang [8]* Inter $\mathcal{C}^\dagger$	Siva and Xiang [8]* Intra + Inter $\mathcal{C}'$	MI-SVM [17]*	Nguyen et al. [5]*	Φ	N	N+Φ
aeroplane	31.1	41.2	**45.4**	37.8	30.7	32.4	**45.4**	38.7
bicycle	18.5	17.7	20.6	17.7	16.5	16.9	20.2	**22.2**
bird	25.2	28.2	**29.7**	26.7	23.0	22.4	29.1	27.6
boat	13.8	13.3	12.2	13.8	14.9	19.9	14.9	**21.0**
bottle	03.3	05.3	04.1	04.9	04.9	04.1	04.5	**06.6**
bus	31.2	35.5	**37.1**	34.4	29.6	31.2	31.7	33.3
car	26.7	33.0	**41.0**	33.7	26.5	34.2	34.2	39.4
cat	42.1	50.1	**53.4**	46.6	35.3	41.8	48.4	46.0
chair	06.7	05.4	06.5	05.4	07.2	07.6	07.6	**08.1**
cow	28.4	30.5	31.9	29.8	23.4	31.2	27.7	**34.8**
diningtable	24.0	16.0	20.5	14.5	20.5	29.0	23.5	**31.5**
dog	33.3	36.1	**40.9**	32.8	32.1	33.0	35.9	38.0
horse	30.7	**38.7**	37.3	34.8	24.4	32.4	34.5	37.6
motorbike	34.7	44.1	**46.5**	41.6	33.1	38.8	39.6	43.3
person	14.3	20.6	22.3	19.9	17.2	21.3	21.9	**23.0**
pottedplant	09.4	11.4	10.2	11.4	12.2	09.4	**14.3**	11.4
sheep	26.0	25.0	27.1	25.0	20.8	**31.3**	25.0	28.1
sofa	25.3	23.6	32.3	23.6	28.8	24.0	29.7	**34.5**
train	42.9	47.9	**49.0**	45.2	40.6	32.2	45.2	43.7
tvmonitor	**10.6**	08.6	09.8	08.6	07.0	09.4	08.2	10.5
Average	23.9	26.6	28.9	25.4	22.4	25.1	27.1	**29.0**
Aeroplane left	42.2	43.8	37.5	42.2	26.6	39.1	**50.0**	39.1
Aeroplane right	38.5	**59.6**	55.8	57.7	25.0	42.3	44.2	50.0
Bicycle left	23.9	26.9	**34.3**	29.9	20.9	19.4	25.4	28.4
Bicycle right	**33.9**	30.7	30.7	19.4	25.8	29.0	27.4	30.6
Boat left	09.4	01.9	09.4	01.9	05.7	**22.6**	09.4	15.1
Boat right	13.8	13.8	12.1	13.8	15.5	**22.4**	10.3	20.7
Bus left	**44.8**	31.0	37.9	37.9	24.1	**44.8**	31.0	31.0
Bus right	29.7	**43.2**	**43.2**	**43.2**	21.6	40.5	29.7	35.1
Horse left	43.9	40.9	48.5	42.4	30.3	34.8	**51.5**	48.5
Horse right	33.9	54.8	51.6	**56.5**	22.6	30.6	45.2	45.2
Motorbike left	46.3	**55.6**	50.0	**55.6**	31.5	42.6	46.3	46.3
Motorbike right	57.5	**66.0**	63.8	63.8	48.9	48.9	48.9	55.3
Average	34.8	39.0	**39.6**	38.7	24.9	34.8	35.0	37.1

*As reported in [8]

Class	[16]	[24]	MI-SVM	Φ	N	N+Φ
boxing	40.7	57.4	20.4	57.4	75.9	**83.3**
clapping	79.4	70.6	61.8	67.6	73.5	**82.4**
handwaving	**93.6**	87.2	85.1	57.4	72.3	78.8
Average	71.2	71.7	55.8	60.8	73.9	**81.5**

Fig. 8. Results using negative mining (N), saliency (Φ), and combined negative mining and saliency ($N + \Phi$) methods

positive nearest neighbour, and *root-normalised* for negative distances. As with the previous section, all scores were linearly mapped onto a $[0, 1]$ range.[4] Table 3 shows all comparisons with the intra-class distance on the VOC2007 and MSR2 data-sets.

[4] The exception being Φ on the MSR2 data-set, which was mapped onto $[0, 0.6]$, see section 3.3.

Table 4. Augmention of existing methods by fusing with our negative mining (N) and saliency (Φ)

Data-set	Intra [8]	Intra [8] + N+Φ	Nguyen et al. [5] *MI-SVM Init. Img.* (One Iteration)	Inter [8] *MI-SVM Init. Avg.* (One Iteration)	Inter [8] + N+Φ	Inter [8]+ Intra [8]	Inter [8]+ Intra [8]+ N+Φ	Intra [16]	Intra [16]+ N+Φ
aeroplane	31.1	**37.8**	27.7	41.2	**46.6**	45.4	**45.8**	31.9	**36.6**
bicycle	18.5	**21.8**	19.3	17.7	**21.0**	20.6	**21.8**	14.8	**19.8**
bird	25.2	**27.0**	20.3	28.2	**29.7**	29.7	**30.9**	24.8	**26.1**
boat	13.8	**17.1**	14.4	13.3	**18.8**	12.2	**20.4**	11.6	**17.7**
bottle	03.3	**04.1**	0**5.7**	05.3	05.3	04.1	**05.3**	**04.5**	04.1
bus	**31.2**	**31.2**	31.7	35.5	**37.6**	37.1	**37.6**	**25.8**	23.7
car	26.7	**39.8**	29.0	33.0	**40.7**	**41.0**	40.8	21.5	**22.9**
cat	42.1	**48.7**	32.3	**50.1**	**50.1**	**53.4**	51.6	44.5	**45.4**
chair	06.7	**08.5**	**07.4**	05.4	06.7	06.5	**07.0**	06.1	**06.5**
cow	28.4	**31.9**	24.1	**30.5**	29.8	**31.9**	29.8	27.7	**28.4**
diningtable	24.0	**32.0**	15.0	16.0	**26.5**	20.5	**27.5**	18.0	**24.5**
dog	33.3	**39.7**	33.0	36.1	**39.4**	40.9	**41.3**	30.6	**35.6**
horse	30.7	**38.7**	17.1	38.7	**41.8**	37.3	**41.8**	30.3	**33.1**
motorbike	34.7	**44.9**	30.2	44.1	**45.7**	46.5	**47.3**	33.1	**38.4**
person	14.3	**23.3**	14.9	20.6	**23.8**	22.3	**24.1**	13.9	**14.1**
pottedplant	09.4	**11.4**	**12.7**	11.4	12.2	10.2	**12.2**	**10.6**	8.6
sheep	26.0	**27.1**	24.0	25.0	**28.1**	27.1	**28.1**	17.7	**22.9**
sofa	25.3	**34.9**	18.8	23.6	**29.7**	32.3	**32.8**	18.3	**24.5**
train	42.9	**45.2**	39.5	47.9	**48.3**	**49.0**	48.7	36.0	**42.5**
tvmonitor	**10.6**	10.5	08.2	08.6	**09.8**	**09.8**	09.4	05.5	**11.3**
Average	23.9	**28.8**	26.6	21.3	**29.6**	28.9	**30.2**	21.4	**24.3**
Aeroplane left	**42.2**	**42.2**	31.3	**43.8**	37.5	37.5	**45.3**	32.8	**42.2**
Aeroplane right	38.5	**46.2**	26.9	59.6	**61.5**	**55.8**	53.8	28.8	**42.3**
Bicycle left	23.9	**28.4**	25.4	26.9	**28.4**	**34.3**	31.3	**25.4**	23.9
Bicycle right	**33.9**	27.4	27.4	30.7	**37.1**	**30.7**	30.6	27.4	**30.6**
Boat left	09.4	**17.0**	07.5	01.9	**15.1**	09.4	**13.2**	05.7	**15.1**
Boat right	13.8	**15.5**	**15.5**	13.8	12.1	12.1	**15.5**	05.2	**15.5**
Bus left	**44.8**	34.5	24.1	31.0	**34.5**	**37.9**	**37.9**	34.5	**37.9**
Bus right	29.7	**37.8**	24.3	43.2	**51.4**	43.2	**45.9**	18.9	**37.8**
Horse left	43.9	**50.0**	24.2	40.9	**51.5**	48.5	**50.0**	47.0	**48.5**
Horse right	33.9	**41.9**	03.2	**54.8**	53.2	**51.6**	**51.6**	38.7	**45.2**
Motorbike left	**46.3**	**46.3**	31.5	**55.6**	50.0	**50.0**	**50.0**	**46.3**	42.6
Motorbike right	**57.5**	51.1	53.2	**66.0**	57.4	**63.8**	57.4	**61.7**	53.2
Average	34.8	**36.5**	24.6	39.0	**40.8**	39.6	**40.2**	31.0	**36.2**

This leave-one-out measure shows the maximal accuracy of a nearest neighbour classifier on VOC2007 and MSR2 data-sets given the best possible annotations, as such it can be seen as an approximate upper-bound for the quality of any 1-NN algorithm. Despite this, on the full VOC data-set, the classification accuracy of negative mining substantially out-performs a nearest neighbour classifier based on positive distance alone, and is only 0.6% worse than the combined classifier using both positive and negative distances. Similar results can be seen on the MSR2 data-set– negative mining and saliency together out performs all but the combination of intra-distances, negative mining and saliency. On the VOC6x2 data-set the additional annotations of *left* and *right* and the fact that truncated or occluded models are discarded, reduces the possible changes in appearance and makes intra-distances a more useful measure. Still, the combined measure of all cues is only 2% better than negative mining and saliency, and this is a reasonable trade-off for the much weaker annotation requirements.

5 Conclusion

This work has presented a simple, robust, technique based upon negative mining. By itself, this technique out-performs all previous existing techniques on the MSR2 and VOC2007 (all views) data-sets. We have shown how this technique can be readily combined with other approaches to MIL, by fusing it as an additional potential, and that doing so has always lead to a significant improvement in performance over the original method. Our comprehensive experiments have validated our approach. We believe that the conceptual and implementational simplicity of our approach, alongside its state-of-the-art performance will make it a valuable tool for increasing the performance of many more sophisticated MIL approaches in the future.

References

1. Viola, P.A., Jones, M.J.: Rapid object detection using a boosted cascade of simple features. In: CVPR, pp. 511–518 (2001)
2. Felzenszwalb, P., Girshick, R., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. TPAMI 32(9) (2010)
3. Poppe, R.: A survey on vision-based human action recognition. Image and Vision Computing 28(6), 976–990 (2010)
4. Weinland, D., Ronfard, R., Boyer, E.: A survey of vision-based methods for action representation, segmentation and recognition. Computer Vision and Image Understanding 115(2), 224–241 (2011)
5. Nguyen, M.H., Torresani, L., de la Torre, F., Rother, C.: Weakly supervised discriminative localization and classification: a joint learning process. In: ICCV, pp. 1925–1932 (2009)
6. Deselaers, T., Alexe, B., Ferrari, V.: Localizing Objects While Learning Their Appearance. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part IV. LNCS, vol. 6314, pp. 452–466. Springer, Heidelberg (2010)
7. Pandey, M., Lazebnik, S.: Scene recognition and weakly supervised object localization with deformable part-based model. In: ICCV (2011)
8. Siva, P., Xiang, T.: Weakly supervised object detector learning with model drift detection. In: ICCV (2011)
9. Vapnik, V.N.: Statistical Learning Theory. Wiley-Interscience (1998)
10. Deselaers, T., Ferrari, V.: A conditional random field for multiple-instance learning. In: ICML (2010)
11. Maron, O., Lozano-Perez, T.: A framework for multiple-instance learning. In: NIPS (1998)
12. Chen, Y., Bi, J., Wang, J.: Miles: Multiple-instance learning via embedded instance selection. PAMI 28(12), 1931–1947 (2006)
13. Adelson, E.H.: On seeing stuff: the perception of materials by humans and machines. In: SPIE, vol. 4299, pp. 1–12 (2001)
14. Alexe, B., Deselaers, T., Ferrari, V.: What is an object? In: CVPR, pp. 73–80 (2010)
15. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. IJCV 88(2), 303–338 (2010)
16. Siva, P., Xiang, T.: Weakly supervised action detection. In: BMVC (2011)

17. Andrews, S., Tsochantaridis, I., Hofmann, T.: Support vector machines for multiple-instance learning. In: NIPS, pp. 577–584 (2003)
18. Fu, Z., Robles-Kelly, A., Zhou, J.: MILIS: Multiple Instance Learning with Instance Selection. TPAMI (99) (2010)
19. Muja, M., Lowe, D.G.: Fast approximate nearest neighbors with automatic algorithm configuration. In: International Conference on Computer Vision Theory and Application VISSAPP 2009, pp. 331–340. INSTICC Press (2009)
20. Yuan, J., Liu, Z., Wu, Y.: Discriminative subvolume search for efficient action detection. In: CVPR, pp. 2442–2449 (2009)
21. Cao, L., Liu, Z., Huang, T.S.: Cross-data action detection. In: CVPR (2010)
22. Vedaldi, A., Fulkerson, B.: VLFeat: An open and portable library of computer vision algorithms (2008), `http://www.vlfeat.org/`
23. Laptev, I., Lindeberg, T.: Space-time interest points. In: ICCV, Nice, France, pp. 432–439 (2003)
24. Siva, P., Xiang, T.: Action detection in crowds. In: BMVC (2010)