Weakly Supervised Activity Analysis with Spatio-Temporal Localisation

Feng Gu[a,b], Muralikrishna Sridhar[a], Anthony Cohn[a], David Hogg[a], Francisco Flórez-Revuelta[b], Dorothy Monekosso[c], Paolo Remagnino[b]

[a]*School of Computing, University of Leeds, LS2 9JT, UK*
[b]*Faculty of Science, Engineering and Computing, Kingston University, KT1 2EE, UK*
[c]*School of Computing, Creative Technologies & Engineering, Leeds Beckett University, LS1 3HE, UK*

Abstract

In computer vision, an increasing number of weakly annotated videos have become available, due to the fact it is often difficult and time consuming to annotate all the details in the videos collected. Learning methods that analyse human activities in weakly annotated video data have gained great interest in recent years. They are categorised as "weakly supervised learning", and usually form a multi-instance multi-label (MIML) learning problem. In addition to the commonly known difficulties of MIML learning, i.e. ambiguities in instances and labels, a weakly supervised method also has to cope with large data size, high dimensionality, and a large proportion of noisy examples usually found in video data. In this work, we propose a novel learning framework that iteratively optimises over a scalable MIML model and an instance selection process incorporating pairwise spatio-temporal smoothing during training. Such learned knowledge is then generalised to testing via a noise removal process based on the support vector data description algorithm. According to the experiments on three challenging benchmark video datasets, the proposed framework yields a more discriminative MIML model and less noisy training and testing data, and thus improves the system performance. It outperforms the state-of-the-art weakly supervised and even fully supervised approaches in the literature, in terms of annotating and detecting actions of a single person and interactions between a pair of people.

Key words: Human Activity Analysis, Spatio-Temporal Localisation, Weakly Labelled Video Data, Multi-Instance Multi-Label Learning

1. Introduction

The annotation and detection of human activities have become increasingly significant research problems in the field computer vision, due to the growing demand of analysing large quantities of available videos. However, the proliferation of videos is often unmatched by the availability of detailed spatio-temporal annotation of activities in these videos, mainly due to the laborious nature of such an effort. Thus, much of the annotation comes in a weakly labelled form, where several class labels are simultaneously provided for a single data unit, i.e. a video, without any information about the spatio-temporal locations of the activities. As a consequence, two types of ambiguities result from such a weakly labelled annotation, making it hard to directly apply conventional supervised learning techniques. The first ambiguity is in the instance location, wherein spatio-temporal locations of the true instances that may correspond to activities in a video are not known a priori for training. The second ambiguity is in the instance label, multiple class labels may be associated with a video, while the true label of each individual instance in the video is not known a priori. The ambiguities of instance location and instance label constitute a weakly supervised learning problem, known as multi-instance multi-label (MIML) learning.

Numerous MIML learning techniques have emerged and formed a powerful weakly supervised learning framework, which is capable of simultaneously dealing with instance location and instance label ambiguities mentioned above. These techniques have been applied to various image datasets [38, 30, 33], but barely to videos. Similar to the applications to images, we expect that MIML learning would address the instance location ambiguity by generating multiple instances at different spatio-temporal locations in a video in the form of a bag, which is labelled with respect to the class labels given for the entire video. Then it learns to identify the true instances that correspond to real activities in the video. Additionally, MIML learning would resolve the instance label ambiguity by explicitly modelling interclass correlations and tries to learn the true label of each instance that corresponds to one of the activity classes in the video. While investigating a recently introduced MIML model [33] that we found scalable to video datasets, we have observed that during both training and testing, a large number of noisy samples, completely irrelevant to any activities of interest, tend to have an adverse effect on the model's learning ability. This problem has been recently studied on various datasets for training multi-instance learners using

an approach known as instance selection [6, 11].

In this work, we use a scalable MIML model [33] as the base classifier, and incorporate an innovative spatio-temporal smoothing based instance selection process for the purpose of reducing noise in video data. This forms a novel MIML learning framework for annotating and detecting human activities in weakly labelled video data, where the labelling merely provides the presence of activities in each video but not their spatio-temporal locations. Our contributions can be summarised as follows:

1) An instance selection process is introduced to enforce spatio-temporal smoothing at the bag level, along with the instance classification by the base MIML classifier at the instance level, which is formulated as an energy function similar to the one defined in the minimisation problem of Markov random fields (MRF) [17];
2) A two-step optimisation is applied to alternate iteratively between the base MIML classifier and the instance selection process, aimed at minimising the MRF like energy function until it converges, which provides the knowledge to distinguish the prototype instances potentially associated with the classes of interest from the noisy ones;
3) The learned knowledge of instance selection is then generalised to testing via a noise removal process based on the support vector data description (SVDD) algorithm [28], which learns a description of the prototype instances, to identify noisy instances as outliers during testing.

On application to three benchmark video datasets, we have found that the proposed framework significantly improves the performance in terms of the annotation task (to recognise activities and annotate their spatio-temporal locations in a training video) and the detection task (to recognise activities and detect their spatio-temporal locations in a testing video). The results also suggest that it outperforms the original MIML model [33], the state-of-the-art weakly supervised approaches [25, 24, 18], as well as fully supervised methods [4, 29, 22, 32] in the literature across the three datasets.

The paper is organised as follows: Section 2 provides a review of related work for MIML techniques and weakly supervised action detection; Section 3 details the feature representation of video under the weakly supervised setting; Section 4 formulates the proposed framework and introduces the generation of instances and bags in the setting of weakly supervised action detection; Section 5 describes the experiments, such as data and implementation details; Section 6 demonstrates results and analysis of the experiments;

finally Section 7 concludes this work and points out possible future work.

2. Related Work

Multi-instance learning and multi-label learning have evolved as two separate paradigms until recently [39], where the authors proposed two solutions to bridge them for the MIML learning problem. The first solution transforms each bag of instances into a single instance and then performs multi-label learning. The second solution generalises a multi-instance single label learning algorithm to handle multiple labels. Subsequently, Zha et. al [38] proposed an undirected graphical model for image classification, which simultaneously captures both the connections between class labels and regions (instances), and the correlations among the labels in a single formulation. Its learning and inference process relies on an expectation maximisation algorithm and approximation methods, e.g. the contrastive divergence algorithm [13] and Gibbs sampling [12], which tend to be slow for problems with a large number of instances. In [30], the authors proposed an active learning framework for image annotation that first divides the multi-label problem into a set of binary classification problems and then devises a multi-label set kernel to weight each instance for the multi-instance learning. This framework exhibits limitations when applying to complex datasets, due to its combined polynomial complexity of labels and instances respectively. The methods above are merely designed for the recognition objects in images without any localisation of recognised objects. Therefore they are not directly applicable to more complex video data, for annotating and detecting human activities, with spatio-temporal localisation of recognised activities.

Hu et al. [14] proposed a multi-instance learning framework, SMILE-SVM, to handle ambiguities in the locations of single person's actions in videos of complex scenes. The framework however relies on manually annotated rough locations of an action in a video for training, and it assumes that a video at most contains one true instance of one of the action classes. Therefore, it cannot be directly applied to weakly labelled video data that provides the presence of multiple activities without any spatio-temporal localisation in each video. A multi-instance learning approach that optimises intraclass and interclass distances for action annotation and detection is introduced in [25]. The training does not require the manual annotation of rough locations of actions, but it still has the same assumption of one true instance of one of the action classes per video. It is optimised by a genetic algorithm, which is

known for its slow convergence rate [15]. It is then extended and improved for weakly supervised annotation in [24], by focusing on negative mining in multi-instance learning. Neither of the approaches [25, 24] however explicitly models the interclass relationships within each video (or bag), and thus may struggle in cases where multiple action classes are simultaneously presented in a video (i.e. multi-label learning).

A MIML approach [33] is recently introduced for bag level object recognition in images that feature ambiguities in both the instance location and instance label. The approach has been shown highly scalable to the amount of instances, the dimensionality of feature space, and the number of label classes, which is ideal for complex video data. It trains a set of discriminative multi-instance classifiers and models the interclass correlation among labels by finding a low rank weight matrix. This enforces the classifiers to share weights and perform multi-label learning. This approach however is designed for the classification of a bag rather than each individual instance in the bag. It might not be able to distinguish the positive instances associated with the label classes of interest from the rest in a bag, for the purpose of spatial localisation in images or spatio-temporal localisation in videos. The performance could further deteriorate under the influence of noise, e.g. problems with a low signal-to-noise-ratio (SNR), particularly common in video data. As a result, some means of removing noise from the data, especially those instances completely irrelevant to any label classes of interest, would be beneficial. Recent approaches [6, 11] on instance selection represent the target concept using multiple prototypes that are formed and updated iteratively, thereby simultaneously eliminating many noisy samples in each bag. While instance selection has shown to be a promising direction, it has so far been applied only to multi-instance learning problems but not MIML ones, and has not been extended to testing. This leads to the motivation of this work, that is, to develop a MIML learning framework that is capable of reducing noise from both the training and testing data through instance selection. Such a framework will be applied and evaluated on weakly labelled video data, for purpose of annotating and detecting human activities with spatio-temporal localisation.

3. Feature Representation of Videos

This section details the feature representation of videos for the purpose of activity analysis in a weakly supervised setting. Given a video dataset,

the first step is to transform the videos into a machine readable input feature vector format, which can be directly fed into a machine learning algorithm. For MIML algorithms however, an additional step of generating bags and instances in each bag is required. We extend the approach in [25] to generate instances and bags for representing videos, and make it applicable not only to actions of a single person but also interactions between a pair of people.

3.1. Prior for Temporal Localisation

Let $[t_b, t_e] \cap \mathbb{N}$ be the frame span of a video clip, and $t_b \leq t_a \leq t_e$ be an anchor time point. Given a window size $z \in \mathbb{N}$, three temporal windows are created as $[t_a - z, t_a] \cap \mathbb{N}$, $[t_a - z/2, t_a + z/2] \cap \mathbb{N}$ and $[t_a, t_a + z] \cap \mathbb{N}$. This resembles a temporal sliding window, and it provides the temporal prior of an instance.

3.2. Prior for Spatial Localisation

Since we are merely interested in human activities in this work, a state-of-the-art person detector [10] is used to detect bounding boxes of all the people in each frame for an anchor point t_a. For actions of a single person, each person's bounding box at t_a is expanded with respect to the width and height to cover regions of potential arm of leg motions. While for interactions between a pair of people, each pair of person bounding boxes are selected and expanded with respect to the widths and heights, and then the union of the two expanded bounding boxes is treated as the final output. This is used as the spatial prior of an instance.

3.3. Generation of Bags and Instances

A set of XYT cuboids are generated for a video clip based on the temporal and spatial priors. We then use the BoWs approach to compute the feature representation of each cuboid, from local descriptors, such as spatio-temporal interest points (STIP) [16] and improved dense trajectory descriptor (IDT) [31]. Some regions of interest (ROI) are first identified in the spatio-temporal space of a video, i.e. interest points and trajectories. Visual features, e.g. histograms of oriented gradients (HOG), histograms of optical flows (HOF) and motion boundary histograms (MBH), are then extracted from each ROI to represent local appearances and motions. A subset of extracted descriptors are randomly selected from the training data, and they are then fed into a unsupervised clustering method, e.g. K-Means [3], to generate a user defined number of clusters. The generated clusters are called

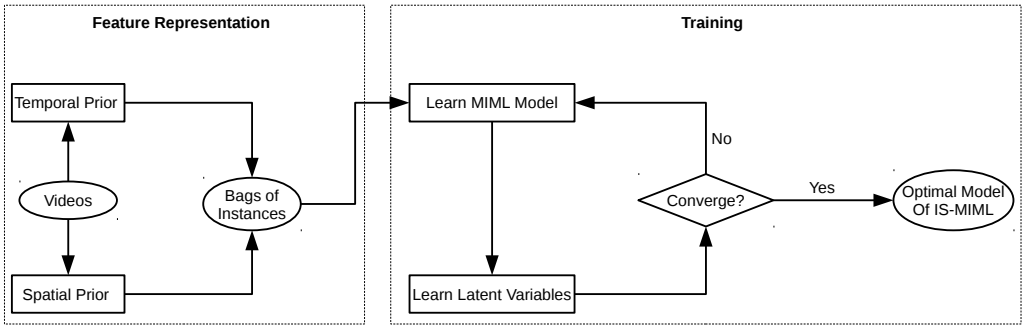


Figure 1: A flow chart of feature representation and training of the proposed methods.

"visual words", and together they form a codebook. All descriptors inside a cuboid are then compared to every visual word in the codebook to compute a high dimensional and sparse histogram feature vector $\mathbf{x} \in \mathbb{N}^D$ ($D \in \mathbb{N}$ is the dimensionality), by counting the numbers of closest clusters of visual words based on some metric, e.g. the Euclidean distance. An instance is thus represented as $(\mathbf{c}, \mathbf{x})$, where $\mathbf{c}$ defines its spatio-temporal cuboid and $\mathbf{x}$ represents its feature vector of appearance and motion. All the instances in a video clip are put into a bag that is labelled with respect to the activity classes presented in the clip. The generated bags of instances are then provided to a training process for learning and inference, as shown in Figure 1.

4. Learning and Inference

Let $(\mathbf{x}_1, \mathsf{y}_1), \ldots, (\mathbf{x}_N, \mathsf{y}_N) \in \mathcal{X} \times \mathcal{Y}$ be a training set, and each example (or video clip) $\mathbf{x}_i$ ($i \in \{1, \ldots, N\}$) is a bag of instances, i.e. $\mathbf{x}_i = \{\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iK_i}\}$ ($K_i = |\mathbf{x}_i|$), where $\mathbf{x}_{ik} \in \mathbb{R}^D$and D is the dimensionality. The spatio-temporal cuboid of an instance $\mathbf{x}_{ik}$ is defined as $\mathbf{c}_{ik} = (x_{ik}, y_{ik}, t_{ik}, w_{ik}, h_{ik}, l_{ik}) \in \mathbb{R}^6$, corresponding to its XYT coordinates, width, height and length. A bag $\mathbf{x}_i$ is assigned a set of labels $\mathsf{y}_i = \{\mathsf{y}_i^1, \ldots, \mathsf{y}_i^M\}$, where $\mathsf{y}_i^j \in \{-1, +1\}$ ($j \in \{1, \ldots, M\}$). A bag is labelled positive with respect to class j, i.e. $\mathsf{y}_i^j = +1$, if at least one of the instances in the bag is positive, while it is labelled negative, i.e. $\mathsf{y}_i^j = -1$, if all the instances are negative. In this work, we assume that the bag is positive for at least one of the label classes, i.e. $\frac{1}{M}\sum_{j=1}^{M} \mathsf{y}_i^j > -1 \quad \forall i$. In the context of activity analysis, a video is labelled positive with respect to an activity class, if at least one instance of the activity is contained in the

video. While the video is labelled negative only if it does not contain any instance of the activity class of interest.

We define two sets of parameters: model parameters $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_M]$, where $\mathbf{w}_j \in \mathbb{R}^D$, is a weight matrix whose columns correspond to the label classes; and instance parameters $\Lambda = [\lambda_1, \ldots, \lambda_N]$, where $\lambda_i \subseteq \{0, 1\}^{K_i}$, is a set of latent variables and λ_{ik} corresponds to the occurrence of an instance $\mathbf{x}_{ik}$ in a training bag $\mathbf{x}_i$. The data are represented as $(\mathbf{x}_1, \lambda_1, \mathsf{y}_1), \ldots, (\mathbf{x}_N, \lambda_N, \mathsf{y}_N)$ by incorporating the latent variables λ_i. The weight matrix parametrises the MIML model, while the latent variable λ_{ik} determines whether an instance $\mathbf{x}_{ik}$ is selected in a bag $\mathbf{x}_i$ for learning, and its value is determined by the cumulative contribution of the instance to the bag being positive with respect to all the label classes. It tends to be one if $\mathbf{x}_{ik}$ is more likely positive for at least one class j, and zero if $\mathbf{x}_{ik}$ is more likely negative for all the classes, that is, to distinguish patterns of any label classes from noise. The objective of instance selection is to remove noisy instances from each bag in the training set, while retaining the positive instances potentially associated with the label classes of interest. It is to minimise the interference from noise and thus reduce the difficulty in discriminating between different classes.

4.1. Energy Function for Learning

In order to learn the two sets of parameters, i.e. model parameters $\mathbf{W}$ and instance parameters Λ, we define an energy function derived from the minimisation problem of MRF [17]. The energy function should be minimised over all the training bags and label classes as

$$\begin{aligned}\mathcal{E}_{\mathbf{W},\Lambda} = & \underbrace{\sum_{i=1}^{N}\sum_{j=1}^{M}\mathcal{L}(\mathsf{y}_i^j, \mathbf{x}_i, \lambda_i, \mathbf{w}_j) + \eta\mathcal{R}(\mathbf{W})}_{\text{unary}} \\ & + \underbrace{\sum_{i=1}^{N}\sum_{ik,ik'\in i}\mathcal{S}(\mathbf{c}_{ik}, \mathbf{c}_{ik'}, \mathbf{x}_{ik}, \mathbf{x}_{ik'}, \lambda_{ik}, \lambda_{ik'})}_{\text{pairwise}} \end{aligned} \tag{1}$$

where the unary term captures a model of appearance and motion with respect to each activity class of interest, while the pairwise term enforces the spatio-temporal smoothness in each bag. On one hand, the unary term is focused on maximising the predictive power of the base MIML classifier in

terms of classifying an instance in each bag as the correct class label. On the other hand, the pairwise term is concentrated on maximising the smoothness of every bag in both time and space, by enumerating all the pairs of neighbouring instances and eliminating those that are similar to each other.

4.1.1. Unary Appearance and Motion Model

The unary term of Equation (1) consists of a loss function for estimating the empirical error of a learned model and a regularisation function for maintaining the model complexity and interclass correlations. It is derived from the MIML model defined in [33], by incorporating the instance parameters Λ. Due to its differentiability and equivalence to maximum margin classifiers [20], the logarithmic loss is chosen and defined as,

$$\begin{aligned}\mathcal{L}(\mathsf{y}_i^j, \mathbf{x}_i, \lambda_i, \mathbf{w}_j) = &- \delta(\mathsf{y}_i^j, 1) \log p(\mathsf{y}_i^j = 1 | \mathbf{x}_i, \lambda_i, \mathbf{w}_j) \\ &- \delta(\mathsf{y}_i^j, -1) \log p(\mathsf{y}_i^j = -1 | \mathbf{x}_i, \lambda_i, \mathbf{w}_j) \end{aligned} \tag{2}$$

where $\delta(a, b) = 1$ if $a = b$ and 0 otherwise. As in logistic regression, the probability of instance $\mathbf{x}_{ik}$ being positive for class j is computed by

$$p(\mathsf{y}_{ik}^j = 1 | \mathbf{x}_{ik}, \lambda_{ik}, \mathbf{w}_j) = \lambda_{ik} \sigma(\mathbf{w}_j \cdot \mathbf{x}_{ik}) = \frac{\lambda_{ik}}{1 + \exp(-\mathbf{w}_j \cdot \mathbf{x}_{ik})} \tag{3}$$

where $\mathbf{w}_j \cdot \mathbf{x}_{ik}$ is the inner product between a weight vector and the feature vector of an instance. An instance $\mathbf{x}_{ik}$ has zero contribution to the bag being positive with respect to any of the label classes, if $\lambda_{ik} = 0$. As a bag is labelled negative only if all instances are negative, we use a noisy-or model to combine probabilities of all the instances in a bag to compute the probability of the bag being negative with respect to class j as

$$\begin{aligned} p(\mathsf{y}_i^j = -1 | \mathbf{x}_i, \lambda_i, \mathbf{w}_j) &= \prod_{k=1}^{K_i} \left(1 - p(\mathsf{y}_{ik}^j = 1 | \mathbf{x}_{ik}, \lambda_{ik}, \mathbf{w}_j)\right) \\ &= \prod_{k=1}^{K_i} \left(1 - \lambda_{ik} \sigma(\mathbf{w}_j \cdot \mathbf{x}_{ik})\right) \end{aligned} \tag{4}$$

Since y_i^j can only have two values, '-1' or '1', the probability of the bag being positive is $p(\mathsf{y}_i^j = 1 | \mathbf{x}_i, \lambda_i, \mathbf{w}_j) = 1 - p(\mathsf{y}_i^j = -1 | \mathbf{x}_i, \lambda_i, \mathbf{w}_j)$.

The regularisation function, scaled by a cost parameter η, is based on the trace-norm defined as

$$\mathcal{R}(\mathbf{W}) = \min_{\mathbf{W}} \|\mathbf{W}\|_\Sigma = \min_{\mathbf{W}=\mathbf{FG}} \frac{1}{2} \left(\|\mathbf{F}\|^2 + \|\mathbf{G}\|^2\right) \tag{5}$$

where $\|\cdot\|$ is the ℓ_2-norm of a matrix. Trace-norm factorises the weight matrix $\mathbf{W}$ into two matrices $\mathbf{F}$ and $\mathbf{G}$, such that $\mathbf{W} = \mathbf{FG}$, to attain a classifier, parametrised by the weight matrix $\mathbf{W}$. The generalization capability of the classifier is thus improved by extracting characteristics that are shared among multiple classes. The matrix $\mathbf{F}$, whose columns define common characteristics, maps the input feature space to an intermediate feature space. While the matrix $\mathbf{G}$, whose columns predict the classes based on the common characteristics, performs classification on the mapped feature space. It aims to derive a low rank matrix that minimises the model complexity while enforcing interclass correlations. This is achieved by minimising the norm of the weight matrix $\mathbf{W}$, which is equivalent to minimising the sum of the norms of matrices $\mathbf{F}$ and $\mathbf{G}$, as defined in Equation (5) [2]. The trace-norm penalty has been proposed for situations where labels are correlated, and it is equivalent to the sum of absolute values of the singular values of the weight matrix [2]. As a result, the trace-norm of the weight matrix $\mathbf{W}$ can be computed by singular value decomposition [8].

4.1.2. Pairwise Spatio-Temporal Smoothing

The pairwise term of Equation (1) is the product of three measures of a pair of instances in a training bag. It is computed over all possible pairs of instances, $K_i(K_i - 1)$ pairs in total, in the bag and is defined as

$$\mathcal{S}(\mathbf{c}_{ik}, \mathbf{c}_{ik'}, \mathbf{x}_{ik}, \mathbf{x}_{ik'}, \lambda_{ik}, \lambda_{ik'}) = \alpha(\mathbf{c}_{ik}, \mathbf{c}_{ik'})\beta(\mathbf{x}_{ik}, \mathbf{x}_{ik'})\varphi(\lambda_{ik}, \lambda_{ik'}) \tag{6}$$

where $\alpha(\mathbf{c}_{ik}, \mathbf{c}_{ik'})$ gives the inverse spatio-temporal overlap between a pair of instances as

$$\alpha(\mathbf{c}_{ik}, \mathbf{c}_{ik'}) = 1 - \frac{\mathbf{c}_{ik} \cap \mathbf{c}_{ik'}}{\mathbf{c}_{ik} \cup \mathbf{c}_{ik'}} \tag{7}$$

and $\beta(\mathbf{x}_{ik}, \mathbf{x}_{ik'})$ computes the inverse cosine similarity metric between the feature vectors of a pair of instances as follows

$$\beta(\mathbf{x}_{ik}, \mathbf{x}_{ik'}) = 1 - \frac{\mathbf{x}_{ik} \cdot \mathbf{x}_{ik'}}{\|\mathbf{x}_{ik}\|\|\mathbf{x}_{ik'}\|} \tag{8}$$

and $\varphi(\lambda_{ik}, \lambda_{ik'})$ returns 1 if $\lambda_{ik} \neq \lambda_{ik'}$ and 0 otherwise. The value of $\alpha(\mathbf{c}_{ik}, \mathbf{c}_{ik'})$ is inversely proportional to the spatio-temporal overlap between a pair of instances, and it penalises those with high spatio-temporal overlaps. The value of $\beta(\mathbf{x}_{ik}, \mathbf{x}_{ik'})$ is inversely proportional to the cosine similarity metric between a pair of instances, and it penalises those with high similarities in

terms of their feature vectors. The value of $\varphi(\lambda_{ik}, \lambda_{ik'})$ is binary and determined by the equality between the instance selection latent variables of a pair of instances, and it penalises those with latent variables of the same value. Essentially the pairwise term encourages a pair of instances that are overlapping spatially, temporally or both and similar to each other in terms of appearance and motion, to have the same value of latent variables. It is inspired by [7], aiming at the identification of clusters corresponding to different label classes that are separable spatially, temporally or both, resulting sparse and more discriminative patterns of appearance and motion. This leads not only to a better spatio-temporal localisation of instances, but also to an improvement on the modelling of interclass correlations parametrised by $\mathbf{W}$. As a result, the pairwise term should significantly improve the proposed framework's ability to handle ambiguities in both the instance location and instance label of a given dataset.

4.2. Two Step Optimisation

In order to solve the minimisation problem of Equation (1), we apply a two-step optimisation that alternates between learning the MIML model $\mathbf{W}$ and learning the latent variables Λ in an iterative manner, as displayed in Figure 1. Such a process is inspired by the optimisation of latent SVM [10, 27]. Let $t \in \mathbb{N}$ be the iteration number of the optimisation process. Parameters of the MIML model at iteration t are denoted as $\mathbf{W}^{(t)} = [\mathbf{w}_1^{(t)}, \ldots, \mathbf{w}_M^{(t)}]$, and the latent variables for instance selection of all the bags are $\Lambda^{(t)} = \{\lambda_i^{(t)}, \ldots, \lambda_N^{(t)}\}$.

4.2.1. Step One: Learning the MIML Model

In the first step, the system learns parameters of the discriminative MIML model $\mathbf{W}$, while latent variables Λ are fixed. The optimisation is performed over a set of continuous variables by minimising the following energy function

$$\mathcal{E}_{\mathbf{W}} = \sum_{i=1}^{N} \sum_{j=1}^{M} \mathcal{L}(\mathrm{y}_i^j, \mathbf{x}_i, \lambda_i, \mathbf{w}_j) + \eta \mathcal{R}(\mathbf{W}) \tag{9}$$

The pairwise term is discarded since it is independent of the MIML model parameters and thus considered a constant. The goal of this step is to find a weight matrix $\mathbf{W}^{(t)}$ that minimises $\mathcal{E}_{\mathbf{W}}$. Such an optimisation problem is non-convex and thus hard to obtain a global optimum. However it can be solved by unconstrained optimisation techniques, to get a sufficiently good

generalisation of unseen data [33]. We use limited-memory BFGS (Broyden-Fletcher-Goldfarb-Shanno) [40], due to its fast convergence rate and low computational cost. It does not explicitly compute and store the Hessian matrix of second-order derivatives, making it highly scalable for large data. Details of the first-order derivatives $\frac{\partial \mathcal{E}_{\mathbf{W}}}{\partial \mathbf{w}_j}$ and $\frac{\partial \mathcal{E}_{\mathbf{W}}}{\partial \mathbf{W}}$, required for the optimisation, can be found in [33].

4.2.2. Step Two: Learning the Latent Variables

In the second step, the MIML model parametrised by $\mathbf{W}$ is fixed, and the minimisation is performed over a set of discrete binary variables Λ as

$$\mathcal{E}_{\Lambda} = \sum_{i=1}^{N} \sum_{j=1}^{M} \mathcal{L}(\mathsf{y}_i^j, \mathbf{x}_i, \lambda_i, \mathbf{w}_j) + \sum_{i=1}^{N} \sum_{ik,ik' \in i} \mathcal{S}(\mathbf{c}_{ik}, \mathbf{c}_{ik'}, \mathbf{x}_{ik}, \mathbf{x}_{ik'}, \lambda_{ik}, \lambda_{ik'}) \quad (10)$$

The regularisation function is dropped since it is independent of the latent variables of instance selection and thus considered a constant. The goal of this step is to search for a set of latent variables $\Lambda^{(t)}$ that minimises $\mathcal{E}_{\Lambda}$. Such an optimisation problem can be solved by simulated annealing [26], to obtain a good approximation of the global optimum [1]. The system draws a subset of Λ from each bag via random sampling, while minimising the value of $\mathcal{E}_{\Lambda}$ by flipping the value of $\lambda_{ik}^{(t)}$. It is performed by finding the minimiser of each bag, which results the minimal sum of all the bags, i.e. $\mathcal{E}_{\Lambda}$. The computation for each bag is independent from that of others, and this process can be performed over all the bags in parallel. As a result, the computational complexity of this step is independent of the number of bags, making it highly scalable for large data as well.

4.2.3. Conditions for Convergence

Either step is guaranteed to converge by the convergence properties of limited-memory BFGS and simulated annealing. For the convergence of the two-step optimisation however, we have to enforce an additional monotonic decrease constraint of the energy function $\mathcal{E}_{\mathbf{W},\Lambda}$, that is, the following criterion must hold throughout the optimisation process

$$\mathcal{E}_{\mathbf{W}^{(t)},\Lambda^{(t)}} \geq \mathcal{E}_{\mathbf{W}^{(t+1)},\Lambda^{(t)}} \geq \mathcal{E}_{\mathbf{W}^{(t+1)},\Lambda^{(t+1)}} \quad (11)$$

The first inequality corresponds to the first step of learning the weight matrix, as defined in Equation (9), using the old bags. The second inequality

refers to the second step of updating the latent variables and effectively each bag, as defined in Equation (10), using the updated weight matrix. Due to the constraint of monotonic value decrease of the energy function $\mathcal{E}_{\mathbf{W},\Lambda}$, as defined in Equation (11), the system is guaranteed to converge in a finite number of steps, or it terminates whenever either of the inequalities is violated. The returned solution at convergence may not be the global optimum of Equation (1), but it is the best solution found by the optimisation process. We denote this iterative learning process as instance selection multi-instance multi-label (IS-MIML).

4.3. Generalisation to Testing

The optimisation of (9) during the training learns the knowledge of prototype instances, of which the latent variables are equal to one. The prototype instances can be seen as patterns that are most likely associated with the label classes, and are distinct from the noise in the training data. It is often the case that the proportion of noisy examples is significantly higher than that of the patterns, and the noise is highly varied and thus difficult to learn a model to sufficiently generalise such a variety. Instead we can treat the patterns as the normal data of an outlier detection problem and try to learn a description of the normal class. Any instance that deviates from such a normal description is considered an outlier or noise and subsequently removed. The remaining instances should be more representative of the prototype instances identified in the training set, and thus easier for the learned MIML model to distinguish between the label classes.

Let $(\mathbf{W}^*, \Lambda^*)$ be the best solution obtained from the two-step optimisation. The prototype instances $\mathbf{x}_{ik}^*$ are selected, such that $\lambda_{ik} = 1$. We use the SVDD algorithm [28] to generalise the learned knowledge of the MIML model and instance selection from the training set to the testing set. The SVDD algorithm is known for its efficiency and robustness of describing high dimensional data for the purpose of outlier detection. It learns a model to provide a closed boundary around the prototype instances $\mathbf{x}_{ik}^*$, i.e. a hypersphere, which is defined by centre $\mathbf{a} \in \mathbb{R}^D$ and radius $R > 0$. The learning aims at minimising the volume of the sphere, determined by R^2, with the constraint that the sphere contains all the prototype instances $\mathbf{x}_{ik}^*$. This can be formulated as an optimisation problem as

$$\begin{cases} \min_{R,\mathbf{a},\xi} & R^2 + C\sum_{ik}\xi_{ik} \\ \text{s. t.} & \|\phi(\mathbf{x}_{ik}^*) - \mathbf{a}\| \leq R^2 + \xi_{ik} \end{cases} \tag{12}$$

where $\phi(\cdot)$ is a kernel function mapping the input data to a higher dimensional feature space, C is a user specified regularization parameter, and ξ_{ik} is the slack variable corresponding to each instance $\mathbf{x}^*_{ik}$. The optimisation above can be solved in an equivalent dual form via quadratic programming [23]. Details of the optimisation in dual form can be found in [28]. Let $\tilde{\mathbf{x}}_{ik}$, $\tilde{\mathbf{c}}_{ik}$, and $\tilde{\lambda}_{ik}$ denote a testing instance, its spatio-temporal cuboid and latent variable. A new instance is detected as an outlier if $\|\phi(\tilde{\mathbf{x}}_{ik}) - \mathbf{a}\| > R^2$, and we define $f(\tilde{\mathbf{x}}_{ik}) = \text{sgn}(R^2 - \|\phi(\tilde{\mathbf{x}}_{ik}) - \mathbf{a}\|)$ for computing the instance selection latent variables. Therefore we have $\tilde{\lambda}_{ik} = \delta(1, f(\tilde{\mathbf{x}}_{ik}))$.

In order to incorporate the spatio-temporal information into the instance selection of testing, we apply a pairwise smoothing function similar to Equation (6) and minimise the following energy function

$$\mathcal{E}_{\tilde{\Lambda}} = \sum_{i=1}^{N} \sum_{ik,ik' \in i} \mathcal{S}(\tilde{\mathbf{c}}_{ik}, \tilde{\mathbf{c}}_{ik'}, \tilde{\mathbf{x}}_{ik}, \tilde{\mathbf{x}}_{ik'}, \tilde{\lambda}_{ik}, \tilde{\lambda}_{ik'}) \tag{13}$$

By minimising $\mathcal{E}_{\tilde{\Lambda}}$ via the same simulated annealing optimisation as in Equation (10), we obtain the optimal latent variables $\tilde{\Lambda} = [\tilde{\lambda}_1, \ldots, \tilde{\lambda}_{\tilde{N}}]$, and $\tilde{\lambda}_{ik} \in \tilde{\lambda}_i$. The final detection score of a testing instance therefore is

$$p(\tilde{y}^j_{ik} = 1 | \tilde{\mathbf{x}}_{ik}, \tilde{\lambda}_{ik}, \mathbf{w}^*_j) = \tilde{\lambda}_{ik} \sigma(\mathbf{w}^*_j, \tilde{\mathbf{x}}_{ik}) \tag{14}$$

where $\mathbf{w}^*_j$ corresponds to the jth column of $\mathbf{W}^*$. We refer the framework with both IS-MIML for training and the SVDD for testing as instance selection one-class multi-instance multi-label (ISOC-MIML).

A detection returned by any of the algorithms described above can be denoted as $(\tilde{\mathbf{c}}_{ik}, p(\tilde{y}^j_{ik} = 1 | \tilde{\mathbf{x}}_{ik}, \tilde{\lambda}_{ik}, \mathbf{w}^*_j))$, i.e. a spatio-temporal cuboid and a detection confidence with respect to an activity class j. Similar to object detection with sliding window approach, there often exist multiple detections of an activity class in each video clip. As a result, we apply a standard intra-class Non-Maximum Suppression (NMS) [19] to identify the top detections (if there are any) of an activity class in each video clip.

5. Experiments

The proposed framework is evaluated on two types of tasks, namely "annotation" and "detection". For both tasks, only the video level ground truth (the presence of activities in a video without spatio-temporal localisation)

is used for training in a weakly labelled setting. For the annotation task, a learned model is required to recognise activities in each training video clip, and to annotate them with spatio-temporal localisation. While for the detection task, a learned model has to detect activities in each testing video clip, with spatio-temporal localisation. The evaluation is aimed to verify the proposed framework's effectiveness in terms of instance selection for both training and testing, and thus IS-MIML and ISOC-MIML are compared to the original MIML [33] and a fully supervised method. We transform the MIML formulation in Equation (9) into a single instance multi-label (SIML) model, where each bag contains one instance generated using the spatio-temporal ground truth. Such a model is essentially equivalent to a multi-class logistic regression classifier that employs trace-norm to regularise interclass correlations as in [2]. In addition to the internal comparisons, the proposed framework is also compared to the state-of-the-art weakly supervised approaches for actions of a single person, e.g. Siva et al. [24] for the annotation task and Siva et al. [25, 18] for the detection task, and fully supervised approaches [4, 29, 22, 32] in the literature.

We adopt a well known evaluation procedure for object detection in the PASCAL visual object classes (VOC) challenge [9]. Let $\mathcal{V} : \mathbb{R}^3 \rightarrow \mathbb{R}$ be a function that computes the volume of a cuboid, i.e. $\mathcal{V}(\mathbf{c}) = w \times h \times l$. Let $\mathbf{c}$ be the spatio-temporal cuboid of a detection, and $\mathbf{g}$ be the spatio-temporal cuboid of an annotated activity in the ground truth. A detection is considered true positive if it is assigned the correct activity class, and satisfies the following constraint of spatio-temporal overlapping $\mathcal{V}(\mathbf{c} \cap \mathbf{g})/\mathcal{V}(\mathbf{c} \cup \mathbf{g}) \geq \varepsilon$. We use $\varepsilon = (1/2)^3 = 1/8$ as in [25, 37, 22], to compute the average precision (AP) of each activity class for all the compared methods.

5.1. Benchmark Video Datasets

Experiments are conducted on three benchmark video datasets, namely "MSR2" [37], "UT interaction" [21], and "LIRIS human activities" [32], for internal comparisons between different versions of the proposed framework and external comparisons with other approaches in the literature.

5.1.1. MSR2 Action Dataset

The MSR2 action dataset [37] is an extended version of the Microsoft research action dataset. It consists of 54 video sequences recorded in a crowded environment with lengths around 40 seconds. The video resolution is 320×240 and frame rate is 15 frames per second. Each video se-

Figure 2: Action examples of the MSR2 action dataset, where different action classes may co-occur temporally.

quence consists of multiple actions that may co-occur temporally but not spatially. Temporal co-occurrence of two activities indicates their temporal durations overlap, and thus cannot be temporally segmented. While spatial co-occurrence of two activities implies their locations in the image plane overlap, i.e. occlusions, and thus they cannot be spatially segmented. There are in total 203 action instances of three classes, namely 'hand waving', 'hand clapping', and 'boxing', as shown in Figure 2. All the video sequences contain at least one instance of each action class, i.e. multi-label. Instances and bags are generated as in Section 3, and about 50,000 instances in total are derived. This gives an average SNR of $1 : 250$. A video-wise leave-one-out cross validation (one video for testing and the rest for training) is used to create the training and testing sets.

5.1.2. UT Interaction Dataset

The UT interaction dataset [21] contains 20 video sequences, each of which is about one minute long. Several participants with more than 15 different clothing conditions appear in the videos. The videos are taken with a resolution of 720×480 and a frame rate of 30 frames per second. There are in total 160 interaction instances of six classes: 'hand shaking', 'hugging', 'pushing', 'pointing', 'punching', and 'kicking', which may also temporally but not spatially co-occur, as shown in Figure 3. All the video sequences contain at least one instance of each interaction class, i.e. multi-label. Similarly, instances and bags are generated as in Section 3, which derives over 12,000 instances. This gives an average SNR of $1 : 75$. A video-wise leave-one-out cross validation is also applied to create the training and testing sets.

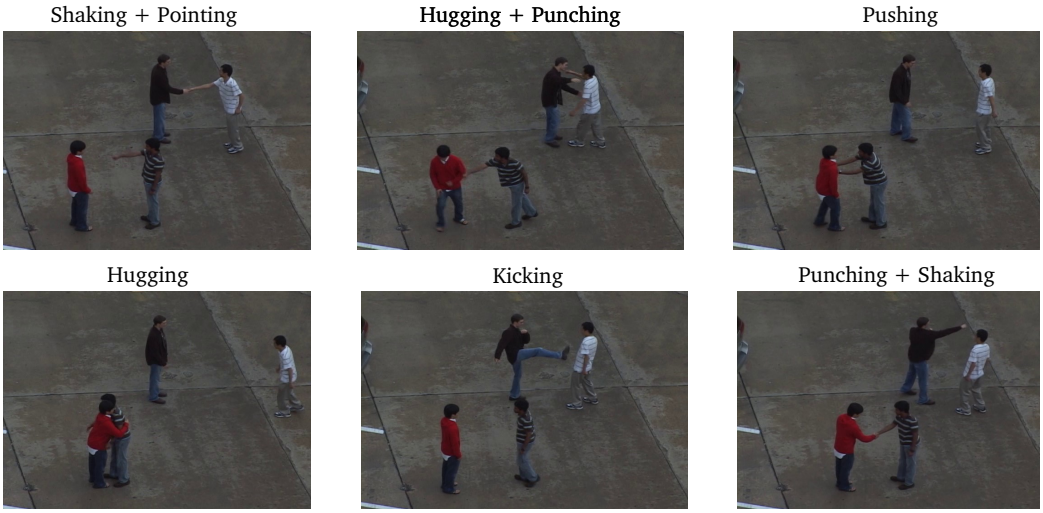


Figure 3: Action examples of the UT interaction dataset, where different interaction classes may co-occur temporally.

5.1.3. LIRIS Human Activities Dataset

The LIRIS human activities dataset [32] is collected for complex and realistic actions and interactions, where each video may contain more than one action or interaction. Overall 21 different actors are involved and activities are shot from various viewpoints and different settings to avoid the possibility of learning activities from background features. There are totally 828 activity instances of 10 classes recorded in 167 video sequences. Among the activities as shown in Figure 4, some of them are interactions between a pair of people, e.g. discussion (DI), give object (GI), hand shaking (HS). Others are characterised as interaction between a person and an object, e.g. put/take box (BO), enter/leave room (EN), try to enter (ET), unlock enter (LO), left baggage (UB), typing and telephone (TE). Simple actions such as walking and running are not considered activities to be detected in this dataset. Two types of cameras were used in the data collection: Primesense/Microsoft Kinect mounted on a mobile robot, capturing RGB images (converted to grey-scale in the publicly available dataset) and 11 bit depth images of a resolution of 640×480, at 25 frames per second; Sony DCR-HC51 camcorder mounted on a tripod, filming RGB images of a resolution of 720×576, at 25 frames per second. Over 90% of the video sequences contained instances of more than one activity class, i.e. multi-label. Similarly,

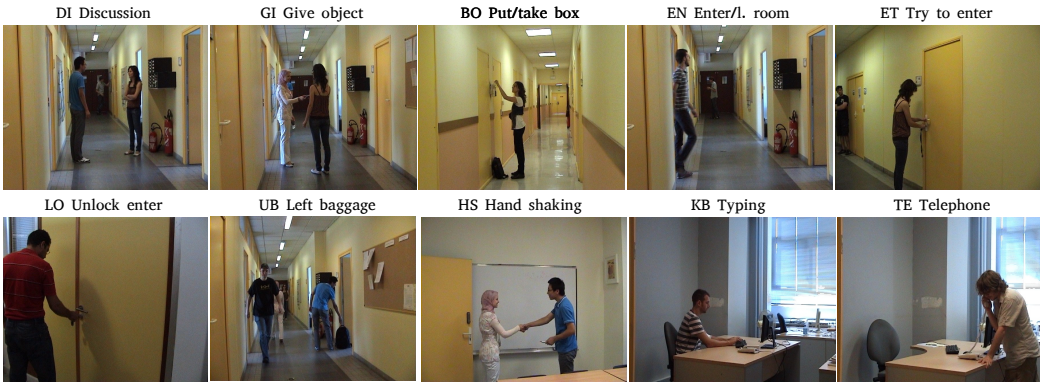


Figure 4: Action examples of the LIRIS human activities dataset, where different action/interaction classes may co-occur temporally.

instances and bags are generated as in Section 3, which derives over 90,000 instances. This gives an average SNR of 1 : 100. A video-wise leave-one-out cross validation is also applied to create the training and testing sets.

The MSR2 dataset exhibits a noticeably low SNR, where the noise is predominantly derived from the crowded environment with lots of background irrelevant motions. The UT interaction dataset is more complicated from an activity analysis perspective, and its SNR is considerably higher. The complexity is resulted from an increased number of activity classes but a decreased number of instances per class. Moreover, activities presented in the UT interaction dataset are interactions between a pair of people, rather than actions of a single person. The proportion of video clips containing multiple activity classes is also significantly higher, which further enhances the difficulty of distinguishing between activity classes. As a result, the MSR2 dataset is mainly used for a baseline evaluation of the proposed framework in comparison to the state-of-the-art weakly supervised approaches [25, 24] applied to the same dataset. While the UT interaction dataset is chosen to test the framework's ability to generalise to more complicated problems. Furthermore, the LIRIS dataset is even more complex than MSR2 and UT interaction datasets. First of all, the size of the dataset and the number of activities classes are significantly larger; secondly some of the activities are more complex, as they involve interactions between a person and an object, e.g. doors, without proper object detectors to identify the semantic labels of these objects, it is very difficult to recognise and detect the corresponding

activities solely based on local descriptors of motion and appearance; Finally the proportion of bags (video clips) that contain more than one activity class and that of noisy instances are significantly greater than the other two datasets, and thus the difficulty of MIML learning has also been enhanced.

5.2. Implementation Details

At the feature representation stage, we use the temporal window size $z = \{30, 60, 90, 120, 150\}$ on all datasets with the anchor point t_a moving at every $z/2$ frames for the temporal prior. Values are chosen with respect to the frame rate and the range of temporal durations of activity classes of all the three datasets. For the spatial prior, each person bounding box returned by the person detector is expanded by 100% to either side in terms of width and 30% upwards in terms of height. Unlike [25], we only remove the cuboids that contain zero STIP or IDT descriptors, to avoid any bias for the generation of instances and bags. For the BoWs method, 500,000 STIP or IDT descriptors are randomly selected from the training data and quantised into 4000 code words, as suggested in [16]. Feature vectors of the derived instances are normalised using the ℓ_2-norm, due to the use of linear kernels in our framework.

The cost parameter η of the MIML model is chosen from $\{2^{-9}, \ldots, 2^{6}\} \cup \{0\}$[1], which yields the highest performance on a validation set (a randomly selected subset of the training set in each fold). The initial temperature of the simulated annealing optimisation is set to 1, and it is decreased by 20% of the previous temperature at every iteration. Due to the stochastic nature of the two-step optimisation, each experiment is run for 10 times and the results are then averaged. We use LIBSVM-SVDD-3.1 [5] with a linear kernel and default parameters for the SVDD. The rest of the framework is implemented in MATLAB. For the intraclass NMS, we use one divided by the number of activity classes as the threshold of confidence, i.e. $p(\tilde{y}_{ik}^{j} = 1 | \tilde{\mathbf{x}}_{ik}, \tilde{\lambda}_{ik}, \mathbf{w}_{j}^{*})$, and 1/8 as the threshold of spatio-temporal overlapping.

[1]Note those values are suggested in [33], and when $\eta = 0$ the regularisation function is effectively disabled.

Method	MSR2	UT Interaction	LIRIS
MIML	0.895	0.584	0.469
IS-MIML	0.921	0.615	0.508
ISOC-MIML	**0.948**	**0.667**	**0.532**

Table 1: Comparison of the original MIML, IS-MIML, and ISOC-MIML in all datasets, for their recognition performances.

6. Results and Analysis

6.1. Qualitative Analysis

In order to demonstrate the proposed methods' capability of removing noisy instances, a qualitative analysis of the effectiveness of instance selection is conducted. First, we compare the recognition performance between the original MIML without instance selection, the IS-MIML method with instance selection for training, and ISOC-MIML method with instance selection for both training and testing. Table 1 lists the average precisions of all the classes in the datasets for the compared methods. The original MIML method gives a reasonable performance, which is on par with the state-of-the-art in the literature. However, the IS-MIML significantly outperforms, which can be due to the introduction of instance selection in training. The instance selection process aims to derive an easier and less noisy training set for a learned model to classify. This is mainly contributed by the addition of the pairwise term in Equation (1) for spatio-temporal smoothing, and the iterative two-step optimisation for learning a more discriminative MIML model. The performance is further improved in the ISOC-MIML, which thus indicates the effectiveness of instance selection in testing.

Moreover, we also randomly select frames from each dataset to evaluate the effect of instance selection for spatio-temporal localisation. Figure 5 displays some examples of the detection results, assuming the classes of activities are accurately recognised. The original MIML method (green bounding boxes) produces larger and inaccurate bounding boxes, while both the IS-MIML (yellow bounding boxes) and ISOC-MIML (blue bounding boxes) give more accurate ones, compared to the ground truth (red bounding boxes). The is due to the fact the spatial location of a candidate instance in a video is predicted by the person detector [10]. It is well known, even the best person detector inevitably produces false positives, and additionally, when the

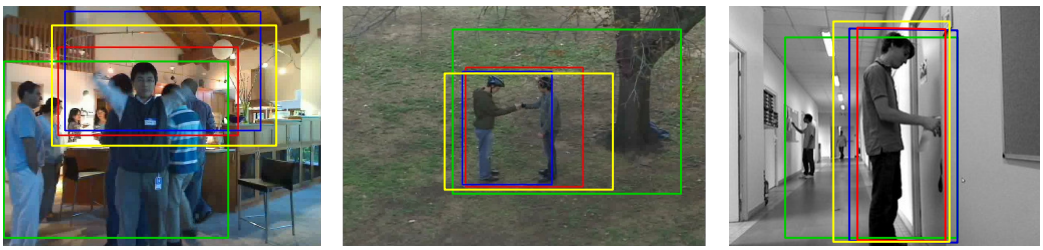


Figure 5: Qualitative comparison of the spatio-temporal location capability between the original MIML, IS-MIML and ISOC-MIML in all datasets, where the red bounding box represents the ground truth, the green one is the output of MIML, the yellow one is the output of IS-MIML, and the blue one is the output of ISOC-MIML.

background is cluttered, detections of people or objects that are completely irrelevant to the activities of interest can be outputted. As a result, without instance selection, the MIML method is unable to remove the noisy instances, which results many instances covering a larger area than an activity of interest does. After the intraclass NMS, the resulting detection would tend to be larger than the ground truth detection, especially in space.

6.2. Quantitative Analysis

As demonstrated in the previous section, the proposed instance selection process has a positive effect on not only the recognition performance, but also the spatio-temporal localisation accuracy. In this section, we conduct a detailed quantitative analysis of the annotation and detection performances of the proposed methods. The annotation performance of each compared method in the MSR2 dataset, UT interaction dataset and LIRIS human activities datasets are listed in Figure 6, Figure 7, and Figure 8 respectively. On the MSR2 dataset, the annotation performance of the IS-MIML/ISOC-MIML method is significantly higher than that of the original MIML, across all the activity classes. IS-MIML/ISOC-MIML also significantly out-performs the state-of-the-art weakly supervised approach [24]. Performances of MIML and IS-MIML/ISOC-MIML are further improved when the feature representation of BoWs is generated from IDT descriptors rather than STIP descriptors. Moreover, in the UT interaction dataset, the IS-MIML/ISOC-MIML method also shows similar performance improvement over the original MIML method, and the improvement gained by switching from STIP descriptors to IDT descriptors in BoWs. Finally, in the LIRIS dataset, the IS-MIML/ISOC-MIML method significantly outperforms the original MIML method, and the

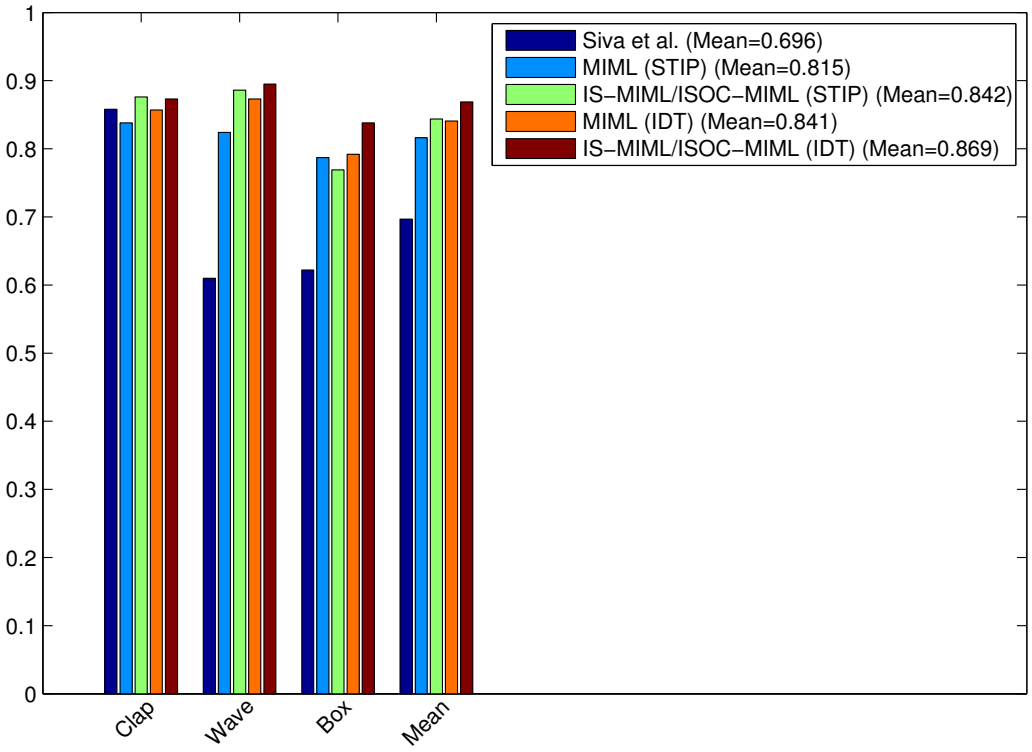


Figure 6: Comparison of the original MIML, the method in Siva et al. [24], and IS-MIML/ISOC-MIML methods on the MSR2 dataset, for the annotation task.

IDT descriptor based BoWs representation produces better results than that with the STIP descriptor.

The annotation results above show that the proposed framework effectively removes noise from the training data, which leads to a more discriminative model compared to the original MIML formulation [33]. The removal of noise also reduces the difficulty of annotation task by eliminating a large number of potential false positives. In comparison to the method in [24], our framework also benefits from explicitly modelling the interclass correlation in each training bag. In addition, the IDT descriptor certainly extracts richer and more robust feature representation than the STIP descriptor for the problem in question, with some sacrifice in computational complexity.

Figure 9 illustrates a comparison of the original MIML, IS-MIML, ISOC-MIML and fully supervised SIML methods using the STIP descriptor or IDT descriptor on the MSR2 dataset, Figure 10 shows the comparison on the UT

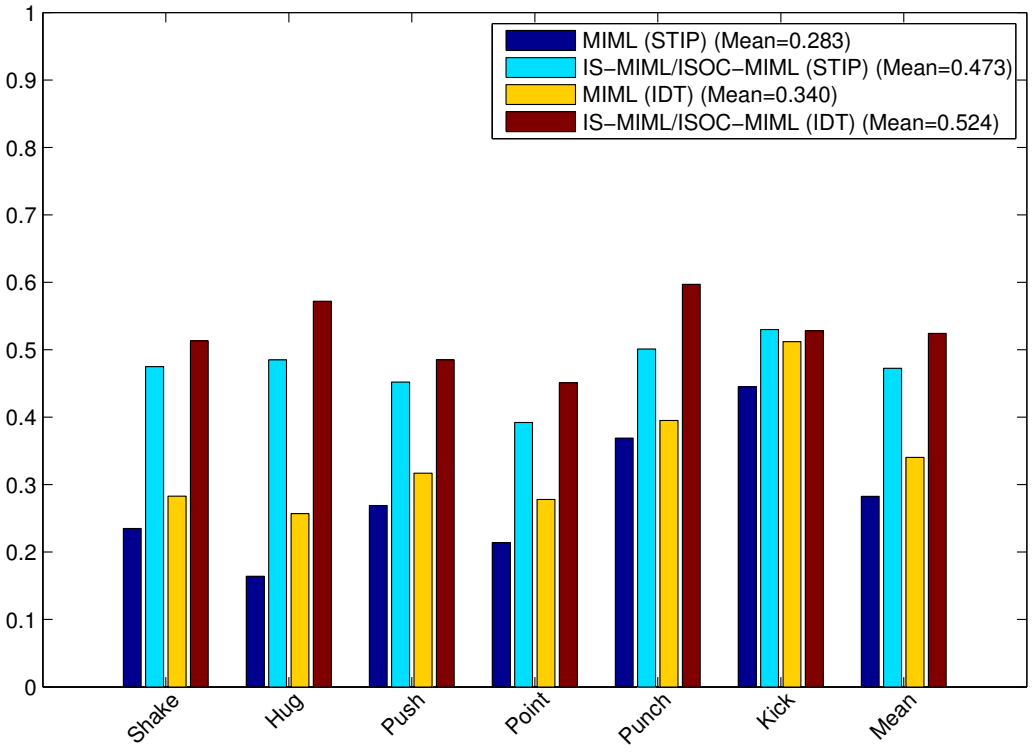


Figure 7: Comparison of the original MIML and IS-MIML/ISOC-MIML methods on the UT interaction dataset, for the annotation task.

interaction dataset, and Figure 11 displays the comparison in the LIRIS dataset. IS-MIML yields a significantly higher detection performance than the original MIML on both datasets. The ISOC-MIML method significantly out-performs IS-MIML, producing overall better performance than the fully supervised SIML method. In addition, similar to what has been observed in the annotation results, the methods using the IDT descriptor produce significantly better results than those using the STIP descriptor. Table 2 lists a comparison between our methods and the state-of-the-art methods in the literature in the MSR2 dataset, Table 3 shows the comparison in the UT interaction dataset, and Table 4 shows the comparison in the LIRIS dataset. The ISOC-MIML method with the IDT descriptor in particular, significantly outperforms the state-of-the-art weakly supervised methods [25, 24, 18] and fully supervised methods [4, 29, 22, 32], across all three datasets.

The detection results above demonstrate the advantage of having instance

Method	Supervision	Avg. Precision
Siva et al. [25]	Full	0.665
Siva et al. [25]	Weak	0.594
Cao et al. [4]	Full (Cross Dataset)	0.170
Tian et al. [29]	Full (Cross Dataset)	0.358
Mosabbeb et al. [18]	Weak	0.701
IS-MIML (STIP)	Weak	0.795
IS-MIML (IDT)	Weak	0.825
ISOC-MIML (STIP)	Weak	0.858
ISOC-MIML (IDT)	Weak	**0.880**

Table 2: Comparison between our methods (ISOC-MIML) and the state-of-the-art methods in the literature for detection task in the MSR2 dataset, where the cross dataset methods are trained on the KTH dataset.

Method	Supervision	Avg. Precision
Shao and Jones [22]	Full	0.375
Yu et al. [35]	Full	0.067
Yu et al [36]	Full	0.299
IS-MIML (STIP)	Weak	0.322
IS-MIML (IDT)	Weak	0.444
ISOC-MIML (STIP)	Weak	0.402
ISOC-MIML (IDT)	Weak	**0.517**

Table 3: Comparison between our methods (ISOC-MIML) and the state-of-the-art methods in the literature for detection task in the UT interaction dataset.

Method	Supervision	Avg. Precision
No. 49 [32]	Full	0.440
NO. A [32]	Full	0.220
No. B [32]	Full	0.470
IS-MIML (STIP)	Weak	0.391
IS-MIML (IDT)	Weak	0.455
ISOC-MIML (STIP)	Weak	0.425
ISOC-MIML (IDT)	Weak	**0.496**

Table 4: Comparison between our methods (ISOC-MIML) and the state-of-the-art methods in the literature for detection task in the LIRIS activities dataset.

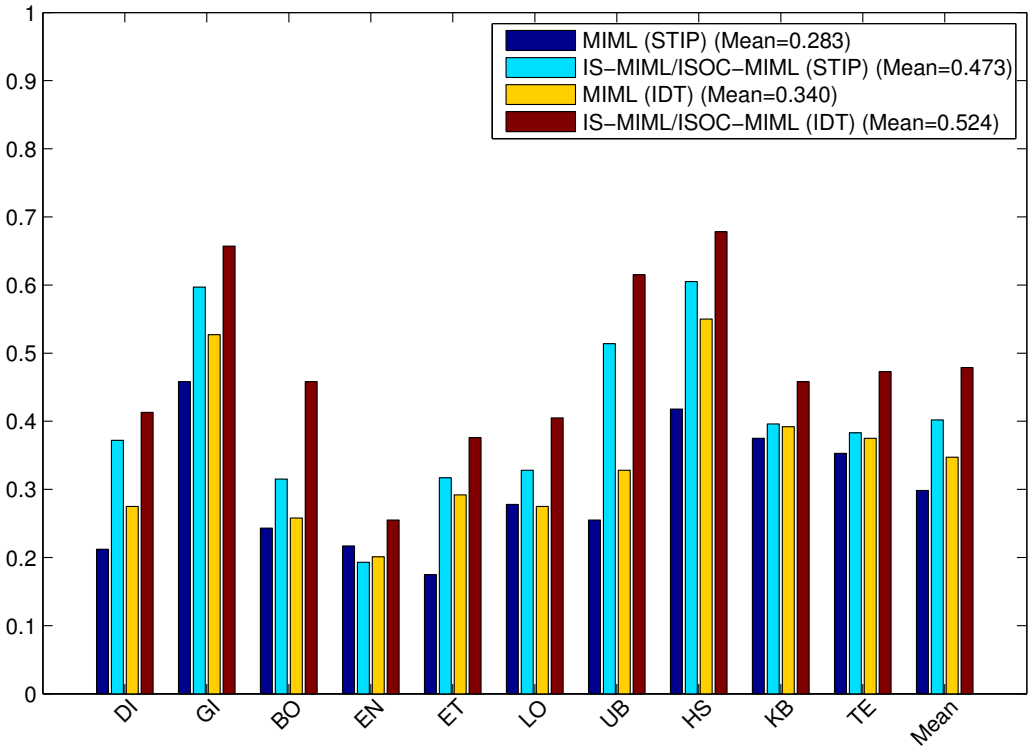
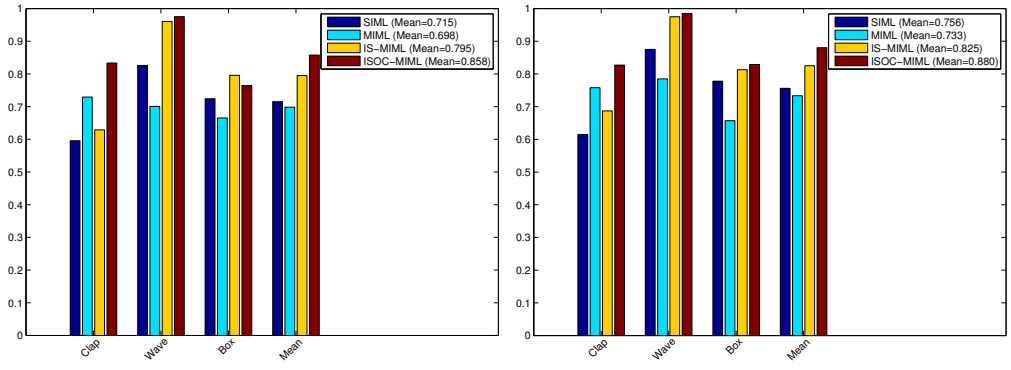


Figure 8: Comparison of the original MIML and IS-MIML/ISOC-MIML methods on the LIRIS human activities dataset, for the annotation task.

selection for training as in th IS-MIML, which leads to a more discriminative MIML model in terms of distinguishing between activity classes. Moreover, the ISOC-MIML method benefits from having instance selection for both training and testing, which gives less noisy testing data in addition to a more discriminative MIML model. As a result, the proposed framework yields a much improved detection performance.

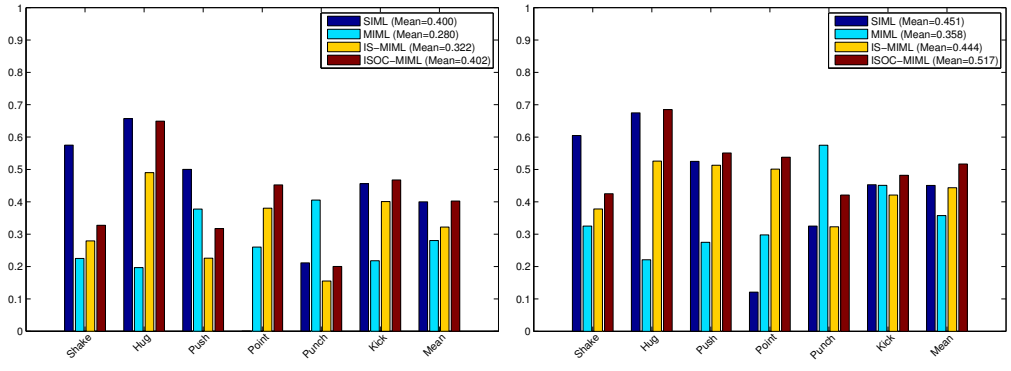
7. Conclusions and Future Work

In this paper, we propose a novel learning framework for annotating and detecting not only actions of a single person but also interactions between a pair of people, in weakly labelled video data. The proposed framework iteratively optimises over a scalable MIML model and an instance selection process incorporating pairwise spatio-temporal smoothing, until the system



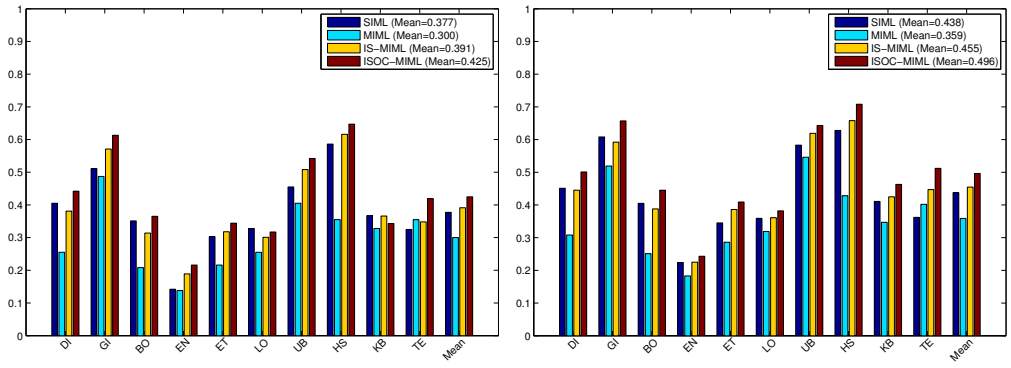
(a) STIP (b) IDT

Figure 9: Comparison of all relevant methods on the MSR2 dataset, for the detection task.



(a) STIP (b) IDT

Figure 10: Comparison of all relevant methods on the UT interaction dataset, for the detection task.



(a) STIP (b) IDT

Figure 11: Comparison of all relevant methods on the LIRIS human activities dataset.

converges. Such learned knowledge is then generalised to the testing phase by a noise removal process based on the SVDD algorithm. According to the experimental results on three benchmark video datasets, the proposed framework effectively removes noise from the training data, which yields a significantly improved annotation performance. It is also proficient at removing noise from the testing data, which leads to a significant improvement on the detection performance. Our framework produces overall better results than a fully supervised method, degenerated from the original MIML model. In addition, it significantly out-performs the state-of-the-art weakly supervised and fully supervised methods in the literature.

There are a number of possible directions for future research, for instance the sliding window-based instance representation can be replaced by a more efficient approach, e.g. spatial-temporal branch-and-bound [37]. This however requires a new MIML model that is able to use such an instance representation. We can also try to model the spatio-temporal structures of activities within a cuboid under a MIML setting, e.g. relationships between body parts as in deformable parts models [34].

Acknowledgements

This work started when Feng Gu was at University of Leeds, working for DARPA Minds Eye project VIGIL (W911NF-10-C-0083). It was then extended and improved during his employment at Kingston University, London, for the project “BREATHE—Platform for self-assessment and efficient

management for informal caregivers" (AAL-JP-2012-5-045).

References

[1] E. H. L. Aarts, J. H. M. Korst, and P. J. M. Laarhoven. *Combinatorial Optimisation*, chapter Simulated annealing, pages 91–120. Wiley-Interscience, 1997.

[2] Y. Amit, M. Fink, N. Srebro, and S. Ullman. Uncovering shared structures in multiclass classification. In *International Conference on Machine Learning (ICML)*, 2007.

[3] C. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[4] L. Cao, Z. Liu, and T.S. Huang. Cross-dataset action detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[5] W. C. Chang, C. P. Lee, and C. J. Lin. A revisit to support vector data description (svdd). Technical report, National Taiwan University, 2013.

[6] Y. Chen, J. Bi, and J. Z. Wang. MILES: Multiple-Instance Learning via Embedded Instance Selection. *IEEE Transactions on PAMI*, 28(12):1931–1947, 2006.

[7] D. J. Crandall and D. P. Huttenlocher. Weakly supervised learning of part-based spatial models for visual object recognition. In *Computer Vision–ECCV 2006*, 2006.

[8] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley-Blackwell, 2nd edition, 2000.

[9] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision (IJCV)*, 88(2):303–338, 2010.

[10] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object Detection with Discriminatively Trained Part Based Models. *IEEE Transactions on PAMI*, 32(9):1627–1645, 2010.

[11] Z. H. Fu, Robles-Kelly A., and J. Zhou. MILIS: Multiple Instance Learning with Instance Selection. *IEEE Transactions on PAMI*, 33(5):958–977, 2011.

[12] S. Geman and D. Geman. Stochastic relaxation, gibbs distribution, and the bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 6(6):721–741, 1984.

[13] G. E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computing*, 14(8):1771–1800, 2002.

[14] Y. Hu, L. Cao, F. Lv, S. Yan, Y. Gong, and T. S. Huang. Action detection in complex scenes with spatial and temporal ambiguities. In *Proceedings of the IEEE 12th International Conference on Computer Vision*, pages 128–135, 2009.

[15] M.T. Jensen. Reducing the run-time complexity of multiobjective EAs: The NSGA-II and other algorithms. *Evolutionary Computation, IEEE Transactions on*, 7(5):503–515, 2003.

[16] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2008.

[17] S. Z. Li. *Markov Random Field Modeling in Image Analysis (Advances in Computer Vision and Pattern Recognition)*. Springer, 3rd edition, 2009.

[18] E. A. Mosabbeb, R. Cabral, F. D. I. Torre, and M. Fathy. Multi-label Discriminative Weakly-Supervised Human Activity Recognition and Localization. In *Asian Conference on Computer Vision (ACCV)*, 2014.

[19] M. Nixon. *Feature Extraction & Image Processing for Computer Vision*. Academic Press, 3rd edition, 2012.

[20] S. Rosset, J. Zhu, and T. Hastie. Margin maximizing loss functions. In *Advances in NIPS*, 2004.

[21] M. S. Ryoo and J. K. Aggarwal. Spatio-Temporal Relationship Match: Video Structure Comparison for Recognition of Complex Human Activities. In *IEEE Conference on ICCV*, 2009.

[22] L. Shao and S. Jones. Efficient Search and Localisation of Human Action in Video Databases. *IEEE Transactions on Circuits and Systems for Video Technology*, 24(3):504–512, 2013.

[23] J. Shawe-Taylor and S. Sun. A review of optimization methodologies in support vector machines. *Neurocomputing*, 74(17):3609–3618, 2011.

[24] P. Siva, C. Russell, and T. Xiang. In defence of negative mining for annotating weakly labelled data. In *European Conference on Computer Vision 2012 (ECCV2012)*, 2012.

[25] P. Siva and T. Xiang. Weakly Supervised Action Detection. In *Proceedings of BMVC*, 2011.

[26] S. Sra, S. Nowozin, and S. J. Wright, editors. *Optimization for Machine Learning*. The, 2012.

[27] M. Tan, G. Pan, Y. Wang, Y. Zhang, and Z. Wu. L1-norm latent svm for compact features in object detection. *Neurocomputing*, 139:56–64, 2014.

[28] D. Tax and R. Duin. Support Vector Data Description. *Machine Learning*, 54:45–66, 2004.

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

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

[31] H. Wang and C. Schmid. Action recognition with improved trajectories. In *IEEE International Conference on Computer Vision (ICCV)*, 2013.

[32] C. Wolf, J. Mille, E. Lombardi, O. Celiktutan, M. Jiu, E. Dogan, G. Eren, M. Baccouche, E. Dellandrea, C.-E. Bichot, C. Garcia, and B. Sankur. Evaluation of video activity localizations integrating quality and quantity measurements. *Computer Vision and Image Understanding*, 127:14–30, 2014.

[33] O. Yakhnenko and V. Honavar. Multi-Instance Multi-Label Learning for Image Classification with Large Vocabularies. In *Proceedings of BMVC*, 2011.

[34] Y. Yang and D. Ramanan. Articulated Human Detection with Flexible Mixtures of Parts. *IEEE Transactions on PAMI*, 35(12):2878–2890, 2013.

[35] G. Yu, J. Yuan, and Z. Liu. Action search by example using randomized visual vocabularies. *IEEE Transaction on Image Processing*, 22(1):377–390, 2013.

[36] G. Yu, J. Yuan, and Z Liu. Propagative Hough Voting for Human Activity Detection and Recognition. *IEEE Transactions on Circuit and Systems for Video Technology*, 25(1):87–98, 2015.

[37] J. Yuan, Z. Liu, and Y. Wu. Discriminative Video Pattern Search for Efficient Action Detection. *IEEE Transactions on PAMI*, 33(9):1728–1743, 2011.

[38] Z. J. Zha, X. S. Hua, T. Mei, J. Wang, G. J. Qi, and Z. Wang. Joint multi-label multi-instance learning for image classification. In *IEEE Conference on CVPR*, 2008.

[39] Z. H. Zhou and M. L. Zhang. Multi-Instance Multi-Label Learning with Application to Scene Classification. In *Advances in NIPS*, pages 1609–1616, 2007.

[40] C. Zhu, R. H. Byrd, P. Lu, and J. Nocedal. Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on Mathematical Software (TOMS)*, 23:550–560, 1997.