Walking and Talking: A Bilinear Approach to Multi-Label Action Recognition

Sameh Khamis Larry S. Davis
University of Maryland
{sameh,lsd}@umiacs.umd.edu

Abstract

Action recognition is a fundamental problem in computer vision. However, all the current approaches pose the problem in a multi-class setting, where each actor is modeled as performing a single action at a time. In this work we pose the action recognition as a multi-label problem, i.e.*, an actor can be performing any plausible subset of actions. Determining which subsets of labels can co-occur is typically treated as a separate problem, typically modeled sparsely or fixed apriori to label correlation coefficients. In contrast, we formulate multi-label training and label correlation estimation as a joint max-margin bilinear classification problem. Our joint approach effectively trains discriminative bilinear classifiers that leverage label correlations. To evaluate our approach we relabeled the UCLA Courtyard dataset for the multi-label setting. We demonstrate that our joint model outperforms baselines on the same task and report state-of-the-art per-label accuracies on the dataset.*

1. Introduction

Action recognition research has recently made tremendous strides. In the past few years research has gone beyond the classic single person short video [2, 29] to model action parts, context, object interactions, group activities, and spatio-temporal connections between actors [6, 12, 16, 27, 34]. However, what all these approaches have in common is that they assume that action recognition is a multi-class problem, where only the most probable label for each actor is predicted.

Multi-class classification is a fundamental problem in machine learning. For many approaches, training is performed in a one-vs-all fashion, where instances from one class are set as positive and the rest negative. Test instances are evaluated and assigned to the class with the highest score. This is appropriate for many problems where labels are mutually exclusive. In semantic segmentation, for instance, each pixel is assigned the name of the class it belongs to. Given that each pixel maps to a single object, and assuming the list of classes do not overlap, multi-class classifications is a natural formulation for the problem [30]. However, if the question we are interested in is "What are they doing?" [7], assigning each actor a single label seems unnecessarily limiting.

Consider the sample frame from the Collective Activity dataset [7] in Figure 1. The two actors in the frame are talking while standing in line, two naturally co-occurring actions. The groundtruth labels for both, however, is the single label *queueing*. In the multi-class setting where a classifier is accordingly only allowed a single label to choose, assigning the label *talking* or *waiting* to either actor is an error and a False Positive for the *talking* or the *waiting* classifier. On the other hand, knowing that the labels *talking*, *queueing*, and *waiting* strongly correlate, a multi-label approach would likely assign the three correct labels to both actors. On the other hand, inversely correlated actions like *queueing* and *crossing* are unlikely to be assigned at the same time to an actor. While the dataset strongly motivated our work, it was not a suitable candidate for our experiments because the actors in most videos were performing the same action.

We propose to treat action recognition as a multi-label classification problem. Each actor can be assigned a subset of the power set of action labels. One can pose multi-label classification as multi-class classification with an exponential number of classes, where each subset of the power set is a separate class. This formulation, however, is computationally infeasible. Equally difficult to solve is formulating multi-label classification as structured prediction for a densely connected Markov Random Field (MRF) of labels, where inference is generally intractable, and typically approaches resort to restricting the structure of the MRF to a tree or at least to small tree width. Instead, we extend recent work on multi-label classification with densely correlated labels [13]. However, instead of assuming an apriori known correlation matrix, we formulate both problems - multi-label training and label correlation estimation - as a joint max-margin bilinear optimization problem. This has the advantage that both problems are optimized to jointly minimize an appropriate loss on the training set. Additionally, discriminatively learning both the classifiers and the

Figure 1. The case for multi-label action recognition. People in natural settings perform more than one action at the same time. Our approach takes into account pairwise correlations to ensure assigned action combinations are meaningful.

label correlations is empirically shown to yield classifiers with better performance accuracy. Finally, given the lack of datasets for our task, we relabeled the UCLA Courtyard dataset [1] using the same set of labels, but instead each actor is assigned a subset of labels instead of a single label.

Our main contribution in this paper is tackling action recognition in the multi-label setting. While attributes, inherently multi-label, have been leveraged before in action recognition to describe the action, the human body configuration, or the manipulated objects, the action recognition problem in itself has always been treated as a multi-class problem. To this end, we introduce a bilinear classification approach where we jointly and discriminatively learn both the classifiers and the label correlations, generalizing previous work where the label correlations were considered prior knowledge or estimated offline.

The rest of this paper is organized as follows. The action and activity recognition literature is surveyed in Section 2. We introduce our joint formulation for multi-label training and correlation estimation in Section 3, and we propose an algorithm to efficiently optimize it. We then present the relabeled UCLA Courtyard dataset and our experimental setup, followed by the evaluation of our approach in Section 4. Finally, we conclude and summarize our work in Section 5.

2. Related Work

Early work in action recognition was mostly concerned with single actors in isolated scenes [2, 29]. However, recently a lot of interest was directed towards modeling the complex interactions among observations explicitly. These interactions could be between scenes and actions [22], objects and actions [11, 34], or actions performed by two or more people [7, 18]. High-level and behavioral interactions were modeled using context-free grammars [27], AND-OR graphs [1, 12], dynamic Bayesian networks [33], network flow [6, 16], and probabilistic first-order logic [4, 21, 24]. However, one common assumption remained: action recognition was formulated as a multi-class problem. To the best of our knowledge, we are the first to formulate action recognition in a multi-label setting.

Recent work that uses attributes for action classification is conceptually related to our work. While attribute and multi-label classification share some of the techniques, semantically speaking they are very different problems. Liu *et al*. recognizes actions from videos by describing them with attributes (indoors, torso-twist, *etc*.) [20]. Yao *et al*. use a mixture of parts and attributes to classify actions in still images [35]. These attributes can represent a description of the action itself (indoors, two-handed), the pose needed (twisted torso, bent elbow, crossed legs), or a manipulated object part (bike seat, golf club). Both approaches classify multiple binary attributes, whether in a pre-processing step or as latent variables, to eventually classify a single action performed by one person in the video or image. In contrast, we are concerned with busy scenes where actors can be performing multiple actions simultaneously, and we are interested in automatically understanding these actions and how they correlate. We accordingly represent the actions as a set of binary inter-dependent labels. Additionally, attributes can still be leveraged and have the potential to benefit multi-label action classification, but we leave this to future work. Mosabbeb *et al*. recently proposed a joint ap-

proach for multi-label action recognition and localization from video, but in their case each video has multiple labels for the actions it contains. In contrast, we label each person in each frame with all the likely actions they are performing.

Early approaches for multi-label classification reduced the problem to more common forms. McCallum proposed to view the problem as a multi-class classification problem with 2^L classes, representing the power set for L labels [23]. While extremely competitive, this approach is very computationally limiting. It also relies on the $0/1$ loss and does not model the multi-label loss [13]. Boutell *et al*. also similarily proposed a power set classifier for multi-label scene classification [3], while Hsu *et al*. proposed a regression-based approach to map the label space to a lower dimensional vector space [14]. Elisseeff and Weston modeled the multi-label loss through a ranking solution [10], where more relevant labels are ranked higher than less relevant ones, and Cai and Hofmann used the same framework to model multi-label loss hierarchically on a tree [5].

Taskar proposed a max-margin structured prediction approach that can be applied to multi-label classification [31]. Structured prediction relies on inference during training, and generally exact inference in MRFs is intractable. Rousu *et al*. extended this to modeling hierarchical loss in a structured prediction setting using a tree-structured model [26]. Restricting the model structure to a tree gives rise to many efficient inference approahes. More recently Petterson and Caetano leveraged MRFs with submodular pairwise interactions [25]. Submodularity also makes efficient inference possible through graph cut algorithms. Hariharan *et al*. took a middle approach by assuming a densely populated pairwise correlation matrix is fixed apriori [13]. Their approach generalizes one-vs-all classifiers in a principled way, and they propose efficient specialized optimization algorithms for it. While an apriori fixed correlation matrix can be expected to be given in a one-shot learning setting[17], it does not readily exist in a general multi-label setting. In our work we extend this approach and jointly optimize the multi-label training and discriminatively estimate the label correlations through a bilinear optimization problem, effectively learning the classifiers and the correlation matrix that together minimize the classification loss on the training set.

3. Approach

3.1. Formulation

We formulate multi-label classification in a max-margin framework. We are given N training samples and a set of L labels. Sample i is represented by $\mathbf{x}_i \in \mathbb{R}^D$ and $\mathbf{y}_i \in \{\pm 1\}^L$, which are respectively its associated feature vector of dimensionality D and label vector of dimensionality L. Each label y_{il} is $+1$ if sample i is a positive sample for label l and -1 otherwise. To this end, we optimize the following objective function [1]

$$F \equiv \min_{\mathbf{W},\mathbf{P}} \frac{1}{2}\|\mathbf{W}\|_F^2 + \lambda\frac{1}{2}\|\mathbf{P}-\mathbf{I}_L\|_F^2 + C\sum_i \max_{\boldsymbol{y}} \left[\Delta(\mathbf{y}_i,\boldsymbol{y}) - (\mathbf{y}_i-\boldsymbol{y})^T\mathbf{P}^T\mathbf{W}^T\mathbf{x}_i\right], \tag{1}$$

where the bilinear classification function is represented by $\mathbf{y} = \mathbf{P}^T\mathbf{W}^T\mathbf{x}$. The hinge loss in Equation 1 penalizes the maximum margin violation for each sample under the loss function of interest. In our case, the loss function Δ represents the misclassification cost if one were to predict label $\boldsymbol{y}$ for $\mathbf{x}_i$ when the true label is $\mathbf{y}_i$.

Hariharan *et al*. introduced a special case of this formulation where they assumed that $\mathbf{P}$ was a known correlation matrix, apriori given or calculated [13]. Their resulting objective is only a function of $\mathbf{W}$. In contrast, we discriminatively learn $\mathbf{P}$ jointly with $\mathbf{W}$ so as to minimize the classification error on the training set. This, in turn, yields stronger bilinear classifiers but complicates the optimization. Our objective function is biconvex (as we will show), and we therefore approach it with an alternating optimization approach.

The formulation in Equation 1 has several advantages. A similar formulation that explicitly models the power set of labels, where the number of classes is 2^L, would equivalently require $N2^L$ constraints, regardless of the loss function used. This proves to be very limiting even for small values of L. On the other hand, Equation 1 under a decomposable loss function has only NL margin constraints. On a different note, modeling the dense pairwise correlations between the labels in a structured prediction framework renders inference, a required step in the optimization process, intractable. A common workaround is to restrict the graph to a tree structure or to impose constraints on the form of correlation (submodularity). In our case the label correlation matrix can be densely specified without negatively affecting the optimization problem.

3.2. Optimization

We approach the problem in Equation 1 using an alternating optimization algorithm. Given a fixed $\mathbf{P}$, we transform F to an SVM-like formulation by substituting $\mathbf{Z} = \mathbf{WP}$ and $\mathbf{R} = \mathbf{P}^T\mathbf{P} \succ 0$ (Positive Semi-Definite) to get the equivalent problem

[1] Our hinge loss is defined similarly to the form commonly used in structured prediction [15, 32] and is therefore slightly different from that in [13].

$$G \equiv \quad \min_{\mathbf{Z}} \frac{1}{2} tr(\mathbf{R}^{-1}\mathbf{Z}^T\mathbf{Z}) + C \sum_i \max_{\boldsymbol{y}} \left[\Delta(\mathbf{y}_i, \boldsymbol{y}) - (\mathbf{y}_i - \boldsymbol{y})^T \mathbf{Z}^T \mathbf{x}_i\right]. \tag{2}$$

The regularization term for $\mathbf{P}$ becomes constant and is dropped. We next assume a decomposable loss function $\Delta(\mathbf{y}_i, \boldsymbol{y}) = \sum_l \delta_l(y_{il}, \boldsymbol{y})$, and then we set the loss function δ_l to the commonly used Hamming loss, inversely weighted by the class frequency for label l to account for class imbalance. For $y_{il} \in \{\pm 1\}$, this simplifies to $\delta_l(y_{il}, -y_{il})$ which we denote by δ_{il} for convenience. Putting everything together, we formulate the objective function equivalently in constrained form

$$\begin{aligned} G \equiv \quad & \min_{\mathbf{Z}, \boldsymbol{\xi}} \frac{1}{2} tr(\mathbf{R}^{-1}\mathbf{Z}^T\mathbf{Z}) + C \sum_i \sum_l \xi_{il} \\ \text{s.t.} \quad & 2y_{il}\mathbf{z}_l^T \mathbf{x}_i \geq \delta_{il} - \xi_{il} \qquad \forall i, l \\ & \xi_{il} \geq 0 \qquad \forall i, l \end{aligned} \tag{3}$$

This is a quadratic matrix programming problem. It can be shown using a Schur complement argument that Equation 3 is convex in $\mathbf{Z}$ if and only if $\mathbf{R} \succ 0$, which is satisfied by definition.

An interesting case arises if we set $\mathbf{P} = \mathbf{I}_L$, where $\mathbf{I}_L$ is the identity matrix of size L. This corresponds to decorrelating the classifiers and recovers the following problem

$$\begin{aligned} G_0 \equiv \quad & \min_{\mathbf{Z}, \boldsymbol{\xi}} \frac{1}{2} \|\mathbf{Z}\|_F^2 + C \sum_i \sum_l \xi_{il} \\ \text{s.t.} \quad & 2y_{il}\mathbf{z}_l^T \mathbf{x}_i \geq \delta_{il} - \xi_{il} \qquad \forall i, l \\ & \xi_{il} \geq 0 \qquad \forall i, l \end{aligned} \tag{4}$$

with in turn is equivalent to L completely independent linear classification subproblems

$$\begin{aligned} G_0 \equiv \quad & \sum_l S_l \\ \text{with } S_l \equiv \quad & \min_{\mathbf{z}_l, \boldsymbol{\xi}} \frac{1}{2} \mathbf{z}^T\mathbf{z} + C \sum_i \xi_i \\ \text{s.t.} \quad & 2y_{il}\mathbf{z}_l^T \mathbf{x}_i \geq \delta_{il} - \xi_i \qquad \forall i \\ & \xi_i \geq 0 \qquad \forall i \end{aligned} \tag{5}$$

Algorithm 1 Cutting plane algorithm for $\mathbf{P}$

INPUT: $\mathbf{V}, \mathbf{Y}, C, \epsilon$
$\mathcal{W} = \emptyset$
repeat
  $\mathcal{P} = \{\mathbf{P} : (p_{ij} = p_{ji} \wedge p_{ij} \geq -1 \wedge p_{ij} \leq 1) \, \forall \, i, j \wedge \frac{2}{N}\mathbf{p}_l^T \sum_i c_{il} y_{il} \mathbf{v}_i \geq \frac{1}{N} \sum_i c_{il}\delta_{il} - \zeta_l \, \forall \, \boldsymbol{c} \in \mathcal{W}\}$
  $\{\mathbf{P}, \boldsymbol{\zeta}\} = \underset{\mathbf{P} \in \mathcal{P}, \boldsymbol{\zeta} > 0}{\operatorname{argmin}} \lambda \frac{1}{2} \|\mathbf{P} - \mathbf{I}_L\|_F^2 + C \sum_l \zeta_l$
  for $l = 1 \ldots L$ **do**
    $c_{il} = \begin{cases} 1 & 2y_{il}\mathbf{p}_l^T \mathbf{v}_i \leq \delta_{il} \\ 0 & \textit{otherwise} \end{cases} \quad \forall i$
  end for
  $\mathcal{W} = \mathcal{W} \cup \{\boldsymbol{c}\}$
until $\max_l (\frac{1}{N} \sum_i c_{il}\delta_{il} - \frac{2}{N}\mathbf{p}_l^T \sum_i c_{il} y_{il} \mathbf{v}_i - \zeta_l) \leq \epsilon$
OUTPUT: $\mathbf{P}$

Algorithm 2 Cutting plane algorithm for $\mathbf{Z}$

INPUT: $\mathbf{X}, \mathbf{Y}, \lambda, C, \epsilon$
$\mathcal{W} = \emptyset$
repeat
  $\mathcal{Z} = \{\mathbf{Z} : \frac{2}{N}\mathbf{z}_l^T \sum_i c_{il} y_{il} \mathbf{x}_i \geq \frac{1}{N} \sum_i c_{il}\delta_{il} - \xi_l \, \forall \, \boldsymbol{c} \in \mathcal{W}\}$
  $\{\mathbf{Z}, \boldsymbol{\xi}\} = \underset{\mathbf{Z} \in \mathcal{Z}, \boldsymbol{\xi} > 0}{\operatorname{argmin}} \frac{1}{2} tr(\mathbf{R}^{-1}\mathbf{Z}^T\mathbf{Z}) + C \sum_l \xi_l$
  for $l = 1 \ldots L$ **do**
    $c_{il} = \begin{cases} 1 & 2y_{il}\mathbf{z}_l^T \mathbf{x}_i \leq \delta_{il} \\ 0 & \textit{otherwise} \end{cases} \quad \forall i$
  end for
  $\mathcal{W} = \mathcal{W} \cup \{\boldsymbol{c}\}$
until $\max_l (\frac{1}{N} \sum_i c_{il}\delta_{il} - \frac{2}{N}\mathbf{z}_l^T \sum_i c_{il} y_{il} \mathbf{x}_i - \zeta_l) \leq \epsilon$
OUTPUT: $\mathbf{Z}$

This simple reduction motivated choosing the identity matrix as the regularization point for $\mathbf{P}$, *i.e.* the regularizer penalizes deviation from it. Similarly, in our optimization procedure, the initial value for $\mathbf{P}$ is $\mathbf{I}_L$. Additionally, this turned out to be an appropriate baseline in our experiments, corresponding to 1-vs-all linear SVM classifiers for the action labels, which is a commonly used benchmark for multi-label methods [3, 9, 36].

We further reduce the number of constraints by employing a one-slack formulation instead [15]. The idea is to replace the N constraints on the hinge loss, one for each of the training samples, with a single constraint on the sum of the hinge losses for all the samples, hence we replace ξ_{il} with one slack variable per label ξ_l. It can be shown that the solution to the one-slack formulation is extremely

sparse and is equivalent to the solution to the original problem if $\xi_l^* = \frac{1}{N}\sum_i \xi_{il}^*$, where $\boldsymbol{\xi}^*$ is the slack vector at the minimum solution [15].

We proceed to solve the one-slack formulation of Equation 3 using a cutting plane approach [32]. At each iteration we find the violated constraints for all the training samples, and we append them to the working set. This algorithm terminates in a number of iterations independent of the output space size [32], and in our experiments we needed fewer than 50 iterations to converge and were faster than the implementation from [13]. The process is detailed in Algorithm 1.

Solving Equation 3 we find $\mathbf{Z}$, and we can then recover $\mathbf{W} = \mathbf{Z}\mathbf{P}^{-1}$. Similarly, given a fixed $\mathbf{W}$, we can turn F to an SVM-like formulation by first transforming the feature vectors to $\mathbf{v}_i = \mathbf{W}^T\mathbf{x}_i$, where each $\mathbf{v}_i$ is of size L, to get the equivalent problem

$$H \equiv \min_{\mathbf{P}} \lambda\frac{1}{2}\|\mathbf{P}-\mathbf{I}_L\|_F^2 + C\sum_i \max_{\boldsymbol{y}} \left[\Delta(\mathbf{y}_i,\boldsymbol{y}) - (\mathbf{y}_i-\boldsymbol{y})^T\mathbf{P}^T\mathbf{v}_i\right]. \tag{6}$$

Under the same decomposable loss function Δ previously introduced, we reformulate the objective function equivalently in constrained form

$$\begin{aligned} H \equiv \quad & \min_{\mathbf{P},\boldsymbol{\zeta}} \lambda\frac{1}{2}\|\mathbf{P}-\mathbf{I}_L\|_F^2 + C\sum_i\sum_l \zeta_{il} \\ \text{s.t.} \quad & 2y_{il}\mathbf{p}_l^T\mathbf{v}_i \geq \delta_{il} - \zeta_{il} \qquad \forall i,l \\ & \zeta_{il} \geq 0 \qquad \forall i,l \end{aligned} \tag{7}$$

Equation 7 is a convex quadratic programming problem. To enforce $\mathbf{P}$ to be a symmetric correlation matrix, we add the constraints $p_{ij} = p_{ji}$, $p_{ij} \geq -1$, and $p_{ij} \leq 1$. We then transform the problem to a one-slack formulation as before, replacing ζ_{il} with one slack variable per label ζ_l. The resulting optimization problem is also solved using a cutting plane algorithm, where we iteratively find the violated constraints for all the training samples, and append them to the working set. The process is detailed in Algorithm 2.

Our alternating optimization approach is illustrated in Algorithm 3. We start by initializing $\mathbf{P}$ to $\mathbf{I}_L$. We then proceed to alternate between fixing $\mathbf{P}$ and solving for $\mathbf{W}$, and then fixing $\mathbf{W}$ and solving for $\mathbf{P}$.

Algorithm 3 Learning Bilinear Multi-Label Classifiers

INPUT: $\mathbf{X}, \mathbf{Y}, \lambda, C, \epsilon, T$
for $t = 1 \ldots T$ **do**
  if $t = 1$ **then**
    Set $\mathbf{P}_t = \mathbf{I}_L$
  else
    Set $\mathbf{v}_i = \mathbf{W}_{t-1}^T\mathbf{x}_i \quad \forall i$
    Calculate $\mathbf{P}_t$ from Algorithm 1
  end if
  Set $\mathbf{R} = \mathbf{P}_t^T\mathbf{P}_t$
  Calculate $\mathbf{Z}_t$ from Algorithm 2
  Set $\mathbf{W}_t = \mathbf{Z}_t\mathbf{P}_t^{-1}$
  if $\max|\mathbf{Z}_t - \mathbf{Z}_{t-1}| < \epsilon$ **then**
    break
  end if
end for
OUTPUT: $\mathbf{P}_t$ and $\mathbf{W}_t$

4. Experiments

4.1. Setup

Given that there are no multi-label action recognition datasets, we set out to relabel an existing datasets for our task. Datasets like KTH [29] and Weizmann [2] feature only a single actor in isolated scenes and are therefore not suitable for a multi-label setting. Similarly, the UT Interaction dataset [28] only features a single action between two actors. On the other hand, we considered the Collective Activity dataset [7]. The dataset features multiple people in different situations, but in most videos all the actors were performing the same action (*e.g.*, dancing), which unfortunately also made it unsuitable for our task.

We set out to relabel the UCLA Courtyard dataset, which features two different bird's eye viewpoints of the same courtyard at the UCLA campus [1]. The dataset features six high resolution videos of many actors in a natural setting performing a variety of actions on both the individual level and the group level. Each actor is annotated by one of 8 orientations, one of 7 poses, and one of 10 individual actions: 1. riding a skateboard, 2. riding a bike, 3. riding a scooter, 4. driving a car, 5. walking, 6. talking, 7. waiting, 8. reading, 9. eating, and 10. sitting. We used the same set of labels for our multi-label experiments. The dataset was evenly split (50-50%) for training and testing, maintaining similar class label distributions for the two halves [1].

Similar to Amer *et al.* [1], we extracted and normalized Histogram of Oriented Gradients (HOG) [8] features around motion-based STIP features and Histogram of Optical Flows (HOF) [19] around KLT tracks from the bounding box of each actor, and therefore the spatial and temporal characteristics were implicitly accounted for through the feature descriptors.

Ultimately the dataset contains over 4.4 million frames, and therefore manually relabeling the entire dataset is very time-consuming. We resorted to bootstrapping the rela-

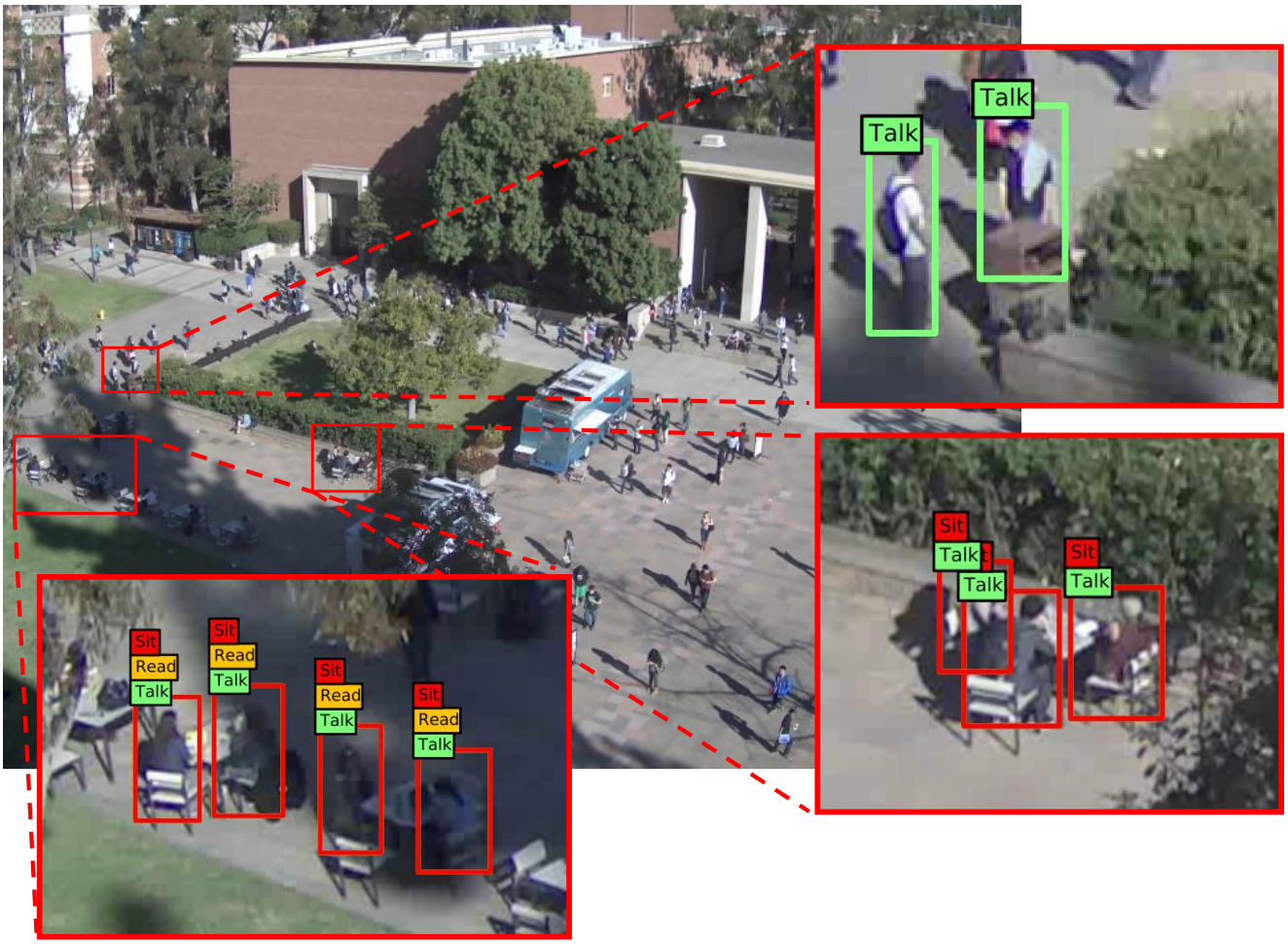

Figure 2. A sample frame from the relabeled UCLA Courtyard dataset.In the resulting labels, 56.9% of all actors are performing two or more actions at the same time and 4.9% are performing three or more actions.

beling process: using the current annotations (pose, orientation, individual action, group action, group orientation, *etc*.), we predict a new set of action labels that include the current action label among others. For instance, a person labeled as *eating* while facing another person, both part of a group labeled as *sitting*, is relabeled as *sitting*, *eating*, and *talking*. We first ran the labels through a large set of similar relabeling rules and then we manually inspected the outcome and optimized the rules to correct any erroneous labels as necessary. This process was repeated a few times to ensure high fidelity for the groundtruth labels. Figure 2 shows a sample frame with multi-label actions. Given the high resolution of the dataset, we zoomed in on a few groups. While the labels for the top group did not change, other groups received additional appropriate action labels. Relabeling was bootstrapped using rules that took into account all the dataset annotations (pose, orientation, individual action, group action, group orientation, *etc*.) to predict new action labels. In the resulting labels, 56.9% of all actors are performing two or more actions at the same time and 4.9% are performing three or more actions.

4.2. Results

Since we initialize the label correlation matrix in our algorithm to the identity matrix $\mathbf{I}_D$, the binary classifiers trained after the first iteration correspond to 1-vs-all linear SVMs trained independently on the same features. This is equivalent to disregarding label correlations and just optimizing Equation 5. Independently training label classifiers in a multi-label setting is an appropriate standard baseline [3, 9, 36], which in our algorithm corresponds to the output after the first iteration. This allows us to evaluate the performance improvement through the iterations by the optimization algorithm. Additionally, we implemented the

multi-label approach of Hariharan *et al*. [13] as a second baseline, where the label correlation matrix is estimated offline from the training data as: $\frac{1}{N}\sum_{i}^{N}\mathbf{y}_i\mathbf{y}_i^T$. Our experiments verify that our approach that discriminatively learns the correlations yields better classification performance.

We report our quantitative results in Table 1. While we are using similar features and data splits to Amer *et al*. [1], we are learning with an entirely different label set, and therefore we cannot directly compare to their results. We include the numbers nonetheless due to the lack of multi-label action recognition datasets and benchmarks. We report the per-class accuracies as well as the mean over all classes. We also report the Hamming loss over all testing samples, which is a common measure for multi-label classification.

As can be seen from the table, our baseline classifier performance is very competitive. We attribute the significant improvement in the mean accuracies to using the weighted hamming loss, in contrast to the Hamming loss (0-1) which optimizes the total classification accuracy. The per-label accuracies for classes like *driving a car*, which looks very unique compared to other classes, is already at 100% after the first iteration. The algorithm converged after 5 iterations of alternating optimization. The improvements, on average, are consistent through the iterations, and more specifically, labels like *reading* and *sitting* received the highest gain through the label correlations, presumably through the correlation with labels like *eating*. Similarly the accuracy for *riding a scooter* also significantly increased, presumably through the correlation with *sitting*. The Hamming loss also decreased through the optimization. It did however slightly increase the last iteration, which again can be attributed to using the weighted hamming loss, which further increased the mean accuracy but slightly sacrificed the total accuracy (1 - Hamming loss).

We also visualize the final label correlation matrix $\mathbf{P}$ calculated by our algorithm in Figure 3. Lighter shades, as seen on the main diagonal, denote positive correlations, and darker shades denote negative (or inverse) correlations. Some of the learned correlations are very intuitive. For example, *walking* and *talking* are likely to co-occur at the same time, which is accurately reflected in the matrix. In contrast, *eating* and *biking* are inversely correlated as expected.

5. Conclusion

We posed action recognition as a multi-label classification problem. Instead of limiting each actor in a natural scene to a single label, we proposed a multi-label setting that is more natural to the problem. Multi-label classification has been either reduced to more common forms, such as multi-class classification, or treated as a Markov Random Field labeling in a structured prediction setting, but both approaches suffer from drawbacks. We instead extended recent work on max-margin multi-label classification to the case where the label correlation matrix is not apriori known, and we posed the multi-label classification and label correlation estimation as a joint problem. We then devised an alternating optimization algorithm to minimize the coupled problem. Finally, given the lack of multi-label action recognition datasets, we relabeled the UCLA Courtyard dataset for our task. We report state-of-the-art results on the dataset using our approach. In future work we plan to investigate integrating group activities into our framework.

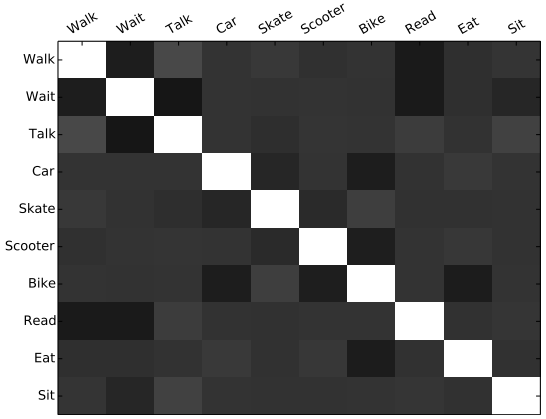

Figure 3. A visualization of the final label correlation matrix $\mathbf{P}$. Intuitively, *walking* and *talking* are positively correlated, while *walking* and *waiting* were unlikely to co-occur in the dataset.

References

[1] M. R. Amer, D. Xie, M. Zhao, S. Todorovic, and S. C. Zhu. Cost-sensitive top-down/bottom-up inference for multiscale activity recognition. In *European Conference on Computer Vision*, 2012. 4322, 4325, 4327

[2] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. In *ICCV*, 2005. 4321, 4322, 4325

[3] M. Boutell, J. Luo, X. Shen, and C. Brown. Learning multi-label scene classification. *PR*, 37(9):17571771, 2004. 4323, 4324, 4326

[4] W. Brendel, S. Todorovic, and A. Fern. Probabilistic event logic for interval-based event recognition. In *CVPR*, 2011. 4322

[5] L. Cai and T. Hofmann. Exploiting known taxonomies in learning overlapping concepts. In *IJCAI*, 2007. 4323

[6] W. Choi and S. Savarese. A unified framework for multi-target tracking and collective activity recognition. In *European Conference on Computer Vision*, 2012. 4321, 4322

[7] W. Choi, K. Shahid, and S. Savarese. What are they doing?: Collective activity classification using spatio-temporal relationship among people. In *VS*, 2009. 4321, 4322, 4325

Approach	Walk	Wait	Talk	Car	Skate	Scooter	Bike	Read	Eat	Sit	Avg	Loss
1-vs-all	72.9	69.7	64.7	100.0	50.2	71.4	51.5	66.2	100.0	83.0	73.0	11.8
two-stage [13]	68.8	73.9	65.3	100.0	54.9	73.1	54.9	76.1	94.6	87.6	74.9	9.8
Our model	70.6	74.7	68.6	100.0	56.8	91.7	58.3	95.2	100.0	87.4	80.3	9.0

Table 1. The quantitative results of our approach. The 1-vs-all baseline, a commonly used baseline for multi-label approaches, corresponds to the resulting classifiers after the first iteration of our model. This is equivalent to optimizing Equation 5 and disregarding label correlations. The second baseline is a two-stage approach that estimates the label correlations separately and non-discriminatively [13].

[8] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005. 4325

[9] K. Dembczynski, W. Waegeman, W. Cheng, and E. Hüllermeier. On label dependence and loss minimization in multi-label classification. *ML*, 88(1-2):5–45, 2012. 4324, 4326

[10] A. Elisseeff and J. Weston. A kernel method for multi-labelled classification. In *NIPS*, 2001. 4323

[11] A. Gupta and L. S. Davis. Objects in action: An approach for combining action understanding and object perception. In *CVPR*, 2007. 4322

[12] A. Gupta, P. Srinivasan, J. Shi, and L. S. Davis. Understanding videos, constructing plots learning a visually grounded storyline model from annotated videos. In *CVPR*, 2009. 4321, 4322

[13] B. Hariharan, L. Zelnik-Manor, S. V. N. Vishwanathan, and M. Varma. Large scale max-margin multi-label classification with priors. In *ICML*, 2010. 4321, 4323, 4325, 4327, 4328

[14] D. Hsu, S. Kakade, J. Langford, and T. Zhang. Multi-label prediction via compressed sensing. In *NIPS*, 2009. 4323

[15] T. Joachims. Training linear SVMs in linear time. In *SIGKIDD*, 2006. 4323, 4324, 4325

[16] S. Khamis, V. I. Morariu, and L. S. Davis. Combining per-frame and per-track cues for multi-person action recognition. In *European Conference on Computer Vision*, 2012. 4321, 4322

[17] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009. 4323

[18] T. Lan, Y. Wang, W. Yang, and G. Mori. Beyond actions: Discriminative models for contextual group activities. In *NIPS*, 2010. 4322

[19] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008. 4325

[20] J. Liu, B. Kuipers, and S. Savarese. Recognizing human actions by attributes. In *CVPR*, 2011. 4322

[21] B. London, S. Khamis, S. H. Bach, B. Huang, L. Getoor, and L. S. Davis. Collective activity detection using hinge-loss markov random fields. In *CVPRW*, 2013. 4322

[22] M. Marszalek, I. Laptev, and C. Schmid. Actions in context. In *CVPR*, 2009. 4322

[23] A. McCallum. Multi-label text classification with a mixture model trained by EM. In *AAAIW*, 1999. 4323

[24] V. I. Morariu and L. S. Davis. Multi-agent event recognition in structured scenarios. In *CVPR*, 2011. 4322

[25] J. Petterson and T. S. Caetano. Submodular multi-label learning. In *NIPS*, 2011. 4323

[26] J. Rousu, C. Saunders, S. Szedmak, and J. ShaweTaylor. Kernel-based learning of hierarchical multilabel classification models. *JMLR*, 7:16011626, 2006. 4323

[27] M. S. Ryoo and J. K. Aggarwal. Stochastic representation and recognition of high-level group activities. *IJCV*, 93(2):183–200, 2010. 4321, 4322

[28] M. S. Ryoo and J. K. Aggarwal. UT-Interaction Dataset, ICPR contest on Semantic Description of Human Activities (SDHA). http://cvrc.ece.utexas.edu/SDHA2010/Human_Interaction.html, 2010. 4325

[29] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local svm approach. In *ICPR*, 2004. 4321, 4322, 4325

[30] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In *ECCV*, 2006. 4321

[31] B. Taskar, C. Guestrin, and D. Koller. Max-margin markov networks. In *NIPS*, 2003. 4323

[32] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. Large margin methods for structured and interdependent output variables. *JMLR*, 6:14531484, 2005. 4323, 4325

[33] T. Xiang and S. Gong. Beyond tracking: modelling activity and understanding behaviour. *IJCV*, 67:21–51, 2006. 4322

[34] B. Yao and L. Fei-Fei. Modeling mutual context of object and human pose in human-object interaction activities. *CVPR*, 2010. 4321, 4322

[35] B. Yao, X. Jiang, A. Khosla, A. L. Lin, L. J. Guibas, and L. Fei-Fei. Human action recognition by learning bases of action attributes and parts. In *ICCV*, 2011. 4322

[36] M. Zhang and Z. Zhou. A review on multi-label learning algorithms. *TKDE*, 26(8):1819–1837, 2013. 4324, 4326