Combining Per-Frame and Per-Track Cues for Multi-Person Action Recognition

Sameh Khamis, Vlad I. Morariu, Larry S. Davis

University of Maryland, College Park
{sameh,morariu,lsd}@umiacs.umd.edu

Abstract. We propose a model to combine per-frame and per-track cues for action recognition. With multiple targets in a scene, our model simultaneously captures the natural harmony of an individual's action in a scene and the flow of actions of an individual in a video sequence, inferring valid tracks in the process. Our motivation is based on the unlikely discordance of an action in a structured scene, both at the track level and the frame level (*e.g.*, a person dancing in a crowd of joggers). While we can utilize sampling approaches for inference in our model, we instead devise a global inference algorithm by decomposing the problem and solving the subproblems exactly and efficiently, recovering a globally optimal joint solution in several cases. Finally, we improve on the state-of-the-art action recognition results for two publicly available datasets.

1 Introduction

We introduce a novel framework for human action recognition from videos. We are motivated by the fact that human actions in a video sequence typically follow a natural structured order, on both a scene level and an individual level. Consider the illustration in Figure 1. The person outlined in the left image is queueing, while the person outlined in the right image is waiting to cross the road. Given the appearance and pose similarity, a classifier might return similar scores for both actions for both people. However, the actions performed by the two people at a later time and the actions of people surrounding them can also provide information for the action inference task. This becomes evident when the person on the right starts crossing and nearby pedestrains start doing the same, while the person on the left stays in the queue and is surrounded by other people waiting in line; at this point, their actions become distinguishable.

Tackling this problem reveals three main challenges; action recognition, identity maintenance, and contextual harmony. We propose a representation that solves all three problems simultaneously and efficiently. A joint solution avoids the incoherencies that arise from solving each problem separately. We initially train a linear SVM on the Action Context (AC) descriptor[1], which explicitly accounts for group actions to recognize an individual's action. We use the normalized classifier scores for the action likelihood potentials. We then train an appearance model for identity association. Our association potentials incorporate both appearance cues and action consistency cues. We also train a scene-action harmony potential, which accounts for how an action fits into the general

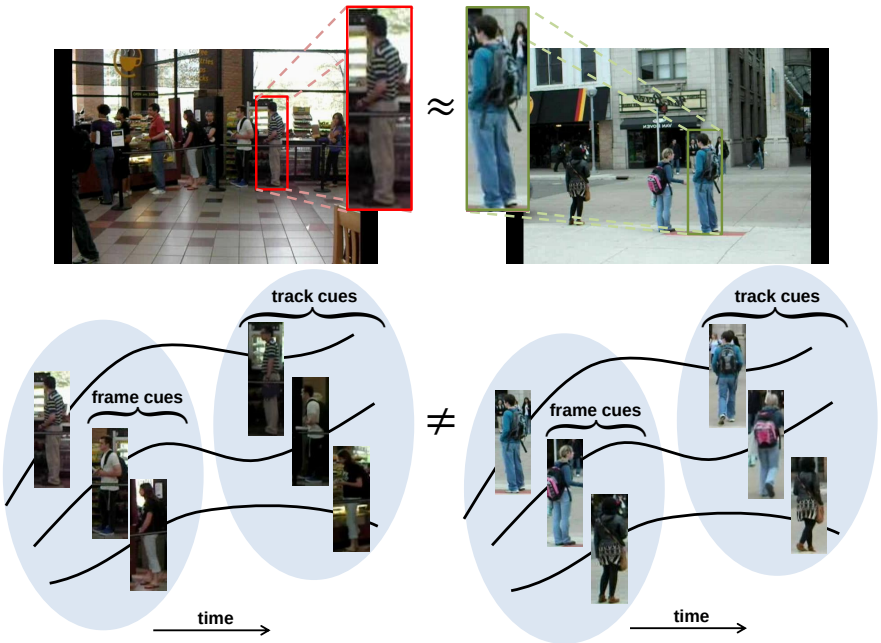

Fig. 1. How per-frame and per-track cues can improve action recognition. While the person outlined on the left is queueing and the person outlined on the right is waiting to cross, a classifier might initially return similar scores for both given the resemblance in their appearance and stance. However, combining tracking and scene harmony, we observe that the person the person on the right is crossing in a street setting, while the person on the left is queueing with other people in line. We present a framework to solve this model with guarantees on global optimality.

setting of the current scene. Our problem can then be naturally represented as a constrained multi-criteria objective function. To obtain a tractable solution, we optimize this function using Dual Decomposition (or Lagrangian Relaxation) by splitting it into two subproblems, both of which are tractable and can be solved exactly and efficiently. Applied to two group action datasets, our approach outperforms state-of-the-art methods.

Our contribution in this work is three-fold:

- We propose a unified model combining per-frame and per-track cues for action recognition, solving identity maintenance in the process.
- We formulate inference as an optimization problem and solve its decompositions exactly and efficiently to recover the joint solution.
- Our action recognition performance improves upon the state-of-the-art results for two publicly available datasets.

The rest of this paper is structured as follows. In Section 2 we survey the action recognition literature and discuss our contribution in its light. We introduce

our approach and focus on the problem formulation in Section 3. We then discuss the system in details in Section 4. In Section 5, we report our quantitative and qualitative results on public datasets. And last, we conclude in Section 6.

2 Related Work

In recent work on action recognition, researchers are explicitly modeling increasingly complex interactions amongst observations, jointly solving multiple previously independent vision problems. Such interactions include those between scenes and actions [2], objects and actions [3, 4] or actions performed by two or more people [5, 6, 1, 7, 8]. More complex high level interactions have also been modeled, e.g., by dynamic Bayesian networks (DBNs) [9], CASE natural language representations [10], Context-Free Grammars (CFGs) [11], AND-OR graphs [12], and probabilistic first-order logic [13, 14].

Most of these approaches require that people or objects are already detected and tracked to incorporate temporal cues [3, 5, 12, 1, 7, 13, 14, 11]. Commonly, tracks are obtained by first detecting people and objects using detectors such as Felzenszwalb *et al.* [15] and then linking the detections to form tracks by various globally-optimal or approximate approaches [16–19]. While performing tracking and activity recognition sequentially simplifies action recognition, mistakes performed during the tracking step cannot be overcome during recognition. Motivated by recent work [8] that obtains improved results by performing identity maintenance and action recognition simultaneously and efficiently, we also solve the identity maintenance problem during action recognition, while modeling additional cues provided by the collective activity in a scene.

Our work is closely related to previous work on modeling collective behavior [5, 1, 6–8]. Choi *et al.* [5] initially modeled collective activity through the construction of a spatio-temporal local (STL) descriptor that relies on an initial 2.5D tracking step to construct histograms of poses (facing left, right, forward, or backward) at binned locations around an anchor person. In later work, Choi *et al.* [7] extend the STL descriptor by using random forests to bin the attribute space and spatio-temporal volume adaptively for better discrimination, then apply an MRF to regularize collective activities in both time and space. Lan *et al.* [1] propose the action context (AC) descriptor, which, unlike the STL descriptor, encodes the actions instead of the poses of people at nearby locations. The AC descriptor stores for each region around a person a k-dimensional response vector obtained from the output of k action classifiers. Instead of relying on local descriptors alone, Lan *et al.* [6] explicitly model group activity by simultaneously modeling the individual actions, their relation to an overall group activity, and their relation to each other. The structure of the person-person interaction graph is inferred as part of the overall inference task. In recent work [8] the AC descriptor is used to perform activity recognition and identity maintenance jointly.

We adopt the AC descriptor and perform joint action recognition and identity maintenance, as in [8], but we also explicitly model the collective activity

in a scene, its effect on individual actions, and its progression over time. Lan et al. [6] model group activities but not the temporal progression of individual actions or group activities. Unlike [6], we do not manually specify a semantically meaningful group activity label, but instead obtain it automatically and use it only to ensure that the activities of people in the same frame are in harmony with each other. While our joint model is complex, we are still able to provide optimality and convergence guarantees without resorting to approximate inference (e.g., sampling) by decomposing the problem into two sub-tasks, a network flow problem and a tree-structured graphical model, both of which can be solved efficiently.

3 Approach

3.1 Overview

Our focus in this work is to improve human action recognition. We assume that humans have already been localized, *e.g.*, with a state-of-the-art multi-part model [15], or with background subtraction if the camera is stationary. Our representation for a detected human figure is based on Histogram of Oriented Gradients (HOG) [20], for which we use the popular implementation from Felzenszwalb *et al.* [15]. We augment our representation with an appearance model for tracking by blurring and subsampling the three color channels of the bounding box in *Lab* color space. We use this representation to train the action and association likelihoods used in our model. We cluster the histograms of actions per-scene for our training data into a set of canonical scene types, which are then used to determine if an action is harmonious with the general setting of the current frame. We present the details of our system in the following sections.

3.2 Formulation

We use i, j, and k to denote the indices of human detections in a video sequence, while a, b, and c are used to denote actions. We also use f to denote frames and s to denote scenes. We define $\mathcal{P}(i)$ to be the set of candidate predecessors for human detection i from prior frames, and similarly $\mathcal{S}(i)$ to be the set of candidate successors of human detection i from subsequent frames. We also define $\mathcal{F}(i)$ to be the frame where human detection i appears. We indicate the action and the identity of a person i by y_i and z_i, respectively, and we indicate the scene type of a frame f by q_f. We can then formulate our model as a cost function over actions, scenes, and identities represented as

$$F(\mathbf{y},\mathbf{q},\mathbf{z}) = \sum_f \sum_s \Bigg[g_s(f) + h_s(f) + \sum_{i \in f} \sum_a \big[u_a(i) + v'_a(i) + w_{sa}(f,i) \big] \; \mathbb{1}(y_i = a) \Bigg] \; \mathbb{1}(q_f = s), \tag{1}$$

where $u_a(i)$ is the classification cost associated with assigning action a to person i, $v'_a(i)$ is the associated tracking cost, and $w_{sa}(f, i)$ is the scene-action harmony cost. $g_s(f)$ denotes the scene prior cost, and $h_s(f)$ denotes the scene consistency cost. Commonly, $\mathbb{1}(.)$ is defined as the indicator function.

We define the classification cost $u_a(i)$ to be the normalized negative classification score of person i performing action a. The details of the classifier training procedure is in Section 4.2.

Since a detection could designate a new person entering the scene, we define our tracking cost as

$$v'_a(i) = \begin{cases} v_{ab}(i,j) & \text{if } \exists j \in \mathcal{P}(i) \text{ s.t. } z_i = z_j,\ y_j = b, \\ \lambda_0 & \text{otherwise,} \end{cases} \tag{2}$$

where $v_{ab}(i, j)$ is the transition cost that links "person i performing action a" to a previously tracked "person j performing action b". If the newly detected person i does not sufficiently match any of the people previously tracked, the model incurs a penalty represented by the tuning parameter λ_0, and a new track is established. We define the transition cost $v_{ab}(i, j)$ as

$$v_{ab}(i,j) = \lambda_d\ d(i,j) - \lambda_c \log(p_{ab}), \tag{3}$$

which is a mixture of an appearance term and an action consistency term. The appearance term measures the similarity between person i and person j with a distance metric $d(i, j)$, and the action consistency term measures the prior probability p_{ab} of a person performing action a followed by action b. The tuning parameters λ_d and λ_c weigh the importance of those two terms. The models for calculating both the appearance distance metric and the action co-occurrences are provided in Section 4.3.

We incorporate scene harmony by modeling a scene using the histogram of the individual actions in that scene. The scene prior cost $g_s(f)$ is calculated as the negative log prior probability p_s of the histogram of actions of scene label s. The scene consistency cost $h_s(f)$ is defined as

$$h_s(f) = \lambda_s \mathbb{1}(q_f \neq q_{f^+}), \tag{4}$$

where f^+ is the next frame. The scene consistency cost is in effect a smoothness prior over scenes in consecutive frames, while the scene-action harmony term $w_{sa}(f, i)$ is defined as

$$w_{sa}(f,i) = -\lambda_h \log(p_{sa}), \tag{5}$$

which models the likelihood p_{sa} of an individual performing action a in a scene labeled s. The tuning parameters λ_s and λ_h weigh the importance of those two terms.

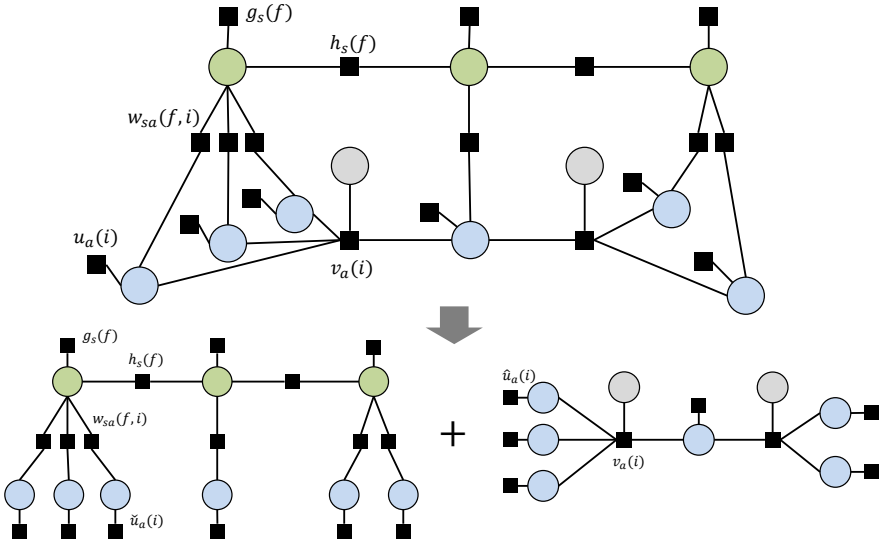

Fig. 2. The representation of our model using factor graph notation. The blue nodes denote the human detections, the green nodes denote the scenes, and the grey nodes denote the identity matching across frames. The potentials presented in Section 3.2 are represented by their associated factor nodes. The decomposition describe in Section 3.3 is shown at the bottom, along with how the potentials are distributed across the subproblems. Refer to the text for more details.

We illustrate our full model in factor graph notation in Figure 2. The blue nodes represent human detections, the green nodes represent scenes, and the grey nodes represent the identity matching between frames. Pairwise cliques tie the scene nodes to all the detections in a specific frame, enforcing a harmonious labeling for the frame, while high-order cliques connect detections across frames to enforce both a valid identity assignment and a valid action-action transition across the tracks. Scene nodes are connected to neighboring scene nodes to discourage abrupt scene label changes.

3.3 Inference

Inference in our model can be formulated as a relaxed integer linear program, but it is more advantageous to leverage the underlying structure of the model. We therefore devise the decomposition illustrated in Figure 2. Maximum-a-posteriori (MAP) estimation in our model can be obtained using a Dual Decomposition optimization scheme [21, 22].

From Equation 1, our model is a function of actions, identities, and scenes. We observe that we can represent the problem via decomposition as

$$\min_{\mathbf{y},\mathbf{q},\mathbf{z}} \; F(\mathbf{y},\mathbf{q},\mathbf{z}) = \min_{\mathbf{y},\mathbf{q},\mathbf{z}} \; \big[F_1(\mathbf{y},\mathbf{q}) + F_2(\mathbf{y},\mathbf{z})\big] \tag{6}$$

where $F_1(.)$ is a function of the actions and scenes in each frame, while $F_2(.)$ is a function of the actions and identities across the tracks. To break the objective function into two parts, we introduce a copy of the *complicating variable* $\mathbf{y}$ for each subproblem and add a consistency (or consensus) constraint to force the two copies to match:

$$\min_{\mathbf{y}_1,\mathbf{y}_2,\mathbf{q},\mathbf{z}} \; \big[F_1(\mathbf{y}_1,\mathbf{q}) + F_2(\mathbf{y}_2,\mathbf{z})\big], \tag{7}$$

$$s.t. \;\; \mathbf{y}_1 = \mathbf{y}_2, \tag{8}$$

We now introduce the the dual variables $\boldsymbol{\nu}$ and form the Lagrangian

$$L(\mathbf{y}_1,\mathbf{y}_2,\mathbf{q},\mathbf{z},\boldsymbol{\nu}) = F_1(\mathbf{y}_1,\mathbf{q}) + F_2(\mathbf{y}_2,\mathbf{z}) + \boldsymbol{\nu}\mathbf{y}_1 - \boldsymbol{\nu}\mathbf{y}_2, \tag{9}$$

which can be separated into two subproblems and yields a lower bound on the optimal solution to the original problem [21]. We then form the dual problem

$$\max_{\boldsymbol{\nu}} \; L(\mathbf{y}_1,\mathbf{y}_2,\mathbf{q},\mathbf{z},\boldsymbol{\nu}) = \tag{10}$$

$$\max_{\boldsymbol{\nu}} \; \Bigg[\underbrace{\min_{\mathbf{y}_1,\mathbf{q}} \big[F_1(\mathbf{y}_1,\mathbf{q}) + \boldsymbol{\nu}\mathbf{y}_1\big]}_{\text{Subproblem 1}} \; + \; \underbrace{\min_{\mathbf{y}_2,\mathbf{z}} \big[F_2(\mathbf{y}_2,\mathbf{z}) - \boldsymbol{\nu}\mathbf{y}_2\big]}_{\text{Subproblem 2}} \Bigg],$$

so that solving the original problem reduces to an iterative process involving the following primal-dual steps:

1. Optimize the two subproblems to obtain the primal variables $\mathbf{y}_1, \mathbf{y}_2, \mathbf{q}, \mathbf{z}$
2. Optimize the dual variables using a subgradient step $\boldsymbol{\nu} = \boldsymbol{\nu} + \eta_t \; (\mathbf{y}_1 - \mathbf{y}_2)$

where η_t is the step size for iteration t [21]. The complicating potentials in our model are the classification cost potentials $u_a(i)$ (see Figure 2) and therefore are distributed evenly across the two subproblems, where each subproblem is then a function of $u_a(i)/2$, for all u and i.

Subproblem 1. The first subproblem is a function of the actions $\mathbf{y}$ and the scenes $\mathbf{q}$ as illustrated on the bottom left of Figure 2. The modified classification cost $\breve{u}_a(i)$ is defined as $u_a(i)/2 + \nu_a(i)$, while the costs $g_s(f)$, $h_s(f)$, and $w_{sa}(f,i)$ are as previously defined. The problem is a tree-structured pairwise graphical model, and hence MAP inference is tractable. We optimize the subproblem exactly and efficiently by maximizing its negative objective function using Max-Product Belief Propagation [23].

Subproblem 2. The second subproblem is a function of the actions $\mathbf{y}$ and the identities $\mathbf{z}$ as illustrated on the bottom right of Figure 2. The high-order cliques in this problem have a special structure; they ensure the validity of the identity assignment between detections, and the consistency of actions across linked detections. While a Belief Propagation algorithm can be formulated for this problems, the time complexity would be pseudo-polynomial [24]. Instead, we use the following integer linear program (ILP) [8]

$$\min_{\mathbf{e},\mathbf{t},\mathbf{x}} \sum_i \sum_a \Big[(\hat{u}_a(i) + \lambda_0) e_a(i) \; + \sum_{j \in \mathcal{P}(i)} \sum_b (\hat{u}_a(i) + v_{ab}(i,j)) t_{ab}(i,j) \Big], \quad (11)$$

$$s.t. \quad e_a(i) \; + \sum_{j \in \mathcal{P}(i)} \sum_b t_{ab}(i,j) \; = \; x_a(i) \; + \sum_{k \in \mathcal{S}(i)} \sum_c t_{ca}(k,i) \quad \forall i, a$$

$$\sum_a \Big[e_a(i) \; + \sum_{j \in \mathcal{P}(i)} \sum_b t_{ab}(i,j) \Big] = 1 \qquad \forall i$$

$$\{\mathbf{e}, \mathbf{t}, \mathbf{x}\} \in \mathbb{B}^n,$$

where variable $e_a(i)$ denotes the entrance of person i into the scene performing action a, while variable $t_{ab}(i,j)$ denotes the transition link of person i performing action a to person j performing action b. Finally, variable $x_a(i)$ denotes person i exiting the scene after performing action a. The entrance, transition, and exit variables are binary indicators. The cost $v_{ab}(i,j)$ is as previously defined, while the modified classification cost $\hat{u}_a(i)$ is defined as $u_a(i)/2 - \nu_a(i)$.

Minimizing the program in Equation 11 is equivalent to inference in the second subproblem from Equation 10. The form of the high-order clique potential between detections of adjacent frames is very sparse. It does not tie the actions of everyone detected in the corresponding frame. It, however, enforces a valid match and thus a valid action transition. The variables $\mathbf{e}$, $\mathbf{t}$, and $\mathbf{x}$ always recover a unique assignment for $\mathbf{y}$ and $\mathbf{z}$. Specifically, if detection i just entered the scene, it will be assigned action $y_i = a$ for which $e_a(i) = 1$ and its identity z_i will be assigned to an unused track number. Otherwise, detection i will be instead linked to a previous detection; in that case, it will be assigned action $y_i = a$ for which $t_{ca}(k,i) = 1$ and the identity will propagate from that previous detection: $z_i = z_k$.

The ILP in Equation 11 represents a network flow problem [8]. In fact, the first constraint of the ILP is the "flow conservation constraint" (or *Kirchoff's Laws*). However, the second constraint, which is refered to as the "explanation constraint", is not typically encountered in the minimum cost flow problem. In this case, it enforces that an action and an identity be assigned to every person detected in the video. The flow of the minimum cost in the network uniquely assigns actions and identities to every detected person in a video sequence.

The contraint matrix of this ILP is totally unimodular [8]. Consequently, we can relax the binary constraint to an interval constraint and still guarantee an integer solution to the linear program. We therefore use a fast interior-point

solver. To improve the inference speed, we only connect people with overlapping bounding boxes in consecutive frames.

Solution Recovery. On convergence, the primal variables $\mathbf{y}_1, \mathbf{y}_2, \mathbf{q}, \mathbf{z}$ and the dual variables $\boldsymbol{\nu}$ are obtained. In the case of an agreement between the two copies $\mathbf{y}_1$ and $\mathbf{y}_2$, the original complicating variable $\mathbf{y}$ is trivially recovered. Otherwise, we recover the best assignment for $\mathbf{y}$ by examining the associated dual variables $\boldsymbol{\nu}$, similar to [22]. The solution is typically tight in 3 iterations, and in several cases the global solution is attained in 6-10 iterations.

4 Learning

4.1 Piecewise Training

Since inference in our model is exact and latent variables are absent, global training approaches become not only possible, but deterministic. However, for practical reasons, we chose to use piecewise training [25]. Piecewise training involves dividing the model into several components, each of which is trained independently. We are motivated by recent theoretical and practical results. Theoretically speaking, piecewise training minimizes an upper bound on the log partition function of the model, which corresponds to maximizing a lower bound on the exact likelihood. In practice, the experiments of [25, 26] show that piecewise training sometimes outperforms global training, even when joint full inference is used. We choose to divide our model training across potentials, and train the groups of potentials independently from each other. The parameters λ_0, λ_c, λ_d, λ_s, and λ_h were manually tuned and ultimately set to 0.25, 0.25, 0.5, 0.1, and 0.25 resepectively for all the experiments.

4.2 Action Potentials

We now describe how we train our action likelihood potentials. We use the AC descriptor from Lan *et al*. [1]. We employ HOG features as the underlying representation. We then train a multi-class linear SVM using *LibLinear* [27]. Next, a bag-of-words style representation for the action descriptor of each person is built. Each person is represented by the associated classifier scores, and the strongest classifier response for every action in a set of defined neighborhood regions in their context.

The descriptor of the i-th person becomes the concatenation of their action scores and context scores. The action scores for person i, given A possible actions, become $\mathbf{F}_i = [s_1(i), s_2(i), \ldots, s_A(i)]$, where $s_a(i)$ is the score of classifying person i to action a. The context score, defined over M neighborhood regions, is

$$\mathbf{C}_i = \left[\max_{j \in \mathcal{N}_1(i)} s_1(j), \ldots, \max_{j \in \mathcal{N}_1(i)} s_A(j), \ldots, \max_{j \in \mathcal{N}_M(i)} s_1(j), \ldots, \max_{j \in \mathcal{N}_M(i)} s_A(j) \right], \tag{12}$$

where $\mathcal{N}_m(i)$ is a list of people in the m-th region in the neighborhood of the i-th person. We use the same "sub-context regions" as [1]. We then run a second-stage classifer on the extracted AC descriptor using the same multi-class linear SVM implementation of *LibLinear* [27]. The classifier scores are negated and then normalized using a softmax function, and finally incorporated as the unary action likelihood potentials $u_a(i)$, which assign action a to person i.

4.3 Association Potentials

To track the identities of the targets in our video sequences, we train identity association potentials and incorporate them in our model. Our association potentials use both appearance and action consistency cues. The appearance cues are trained using the subsampled color channels as features. We train for a Mahalanobis distance matrix M to estimate the similarity between detections across frames. The distance matrix is learned so as to bring detections from the same track closer, and those from different tracks apart [28]. This is formulated as

$$\mathbf{M}^* = \arg\min_{\mathbf{M}} \sum_{\mathcal{T}_k} \Bigg[\sum_{i,j \in \mathcal{T}_k} (\mathbf{f}_i - \mathbf{f}_j)^T \mathbf{M} (\mathbf{f}_i - \mathbf{f}_j) \; - \sum_{\substack{i' \in \mathcal{T}_k \\ j' \notin \mathcal{T}_k}} (\mathbf{f}_{i'} - \mathbf{f}_{j'})^T \mathbf{M} (\mathbf{f}_{i'} - \mathbf{f}_{j'}) \Bigg], \tag{13}$$

where $\mathcal{T}_k$ is the k-th track and $\mathbf{f}_i$ is the feature vector of the i-th person. We solve for M using the fast Large Margin Nearest Neighbor (LMNN) implementation of [29]. The distance between the features of two detected people i and j can then be defined as

$$d(i,j) = (\mathbf{f}_i - \mathbf{f}_j)^T \mathbf{M} (\mathbf{f}_i - \mathbf{f}_j). \tag{14}$$

The action consistency cues are estimated using the groundtruth action labels from the training set. We count pairwise co-occurrences of actions on the same track plus a small additive smoothing parameter α. The counts are normalized into the pairwise co-occurrence probabilities p_{ab} of action pairs a and b.

4.4 Scene Potentials

We cluster the histograms of actions in all the frames of our training set using k-means, where we set $k = 8$ in all of our experiments. The k-means cluster centroids are good representatives of the most likely scenes, and so the centroid histograms are an appropriate approximation for the likelihood of an action given a scene canonical scene types, while the number of points in each cluster is used to approximate the scene prior probability. The form of our scene potentials is similar to the harmony potentials introduced in [30], but our training approach is different.

5 Experiments

5.1 Datasets

We use the group actions dataset from [5] and its augmentation from [7] to evaluate our model. The datasets are appropriate since they have multiple targets in a natural setting, while most action datasets, like KTH [31] or Weizmann [32], have a single person performing a specific action. The original dataset includes 5 action classes: *crossing*, *standing*, *queueing*, *walking*, and *talking*. The augmented dataset includes 6 action classes: *crossing*, *standing*, *queueing*, *talking*, *dancing*, and *jogging*. The *walking* action was removed from the augmented dataset because it is ill-defined [5]. We only use the bounding boxes, the associated actions, and the identities. We did not use any of the 3-D trajectory information.

Our main focus here is action recognition, and tracking is used only to improve the performance in the full model. We evaluate our results similar to [5, 7]. For each dataset, we perform a leave-one-video-out cross-validation scheme. This means that when we classify the actions in one video, we use all the other videos in the dataset for training and validation. Our action potentials are based on [1], which we also compare against to analyze the efficacy of our approach.

5.2 Results

Our confusion matrices for the 5-class and the 6-class datasets are shown in Figure 3. It is clear that removing the *walking* activity improves the classification performance, possibly due to the apparent ambiguity between *walking* and *crossing*. Our average classification accuracy is 72.0% on the former dataset and 85.8% on the latter.

We outperform the state-of-the-art methods on the two datasets, as shown in Table 1. Classification using the AC descriptor that we employ was reported in [1], which we improve upon. The model from [7] employs additional trajectory information, including the 3D location and the pose of every person [7].

We also report qualitative results on the 6-activity dataset in Figure 4. Each row in the figure represents a different video sequence. The first 3 sequences are successful cases where the full model improves the action classification results over either the track cues or the frame cues in isolation, while the final row represents one failure case where the high confidence in the wrong label causes the full model to misclassify the entire frame.

6 Conclusion

We introduced a model that combines tracking cues and scene cues to improve action classification results. The intractability of our model is overcome by a decomposition that leverages its underlying structure. The decomposition yields two subproblems, which we solve exactly and efficiently. We recover the solution to the original problem, which is optimal in several cases. Finally, by combining both cues, we reported action recognition results that outperform the state-of-the-art on two publicly available datasets using the same validation scheme.

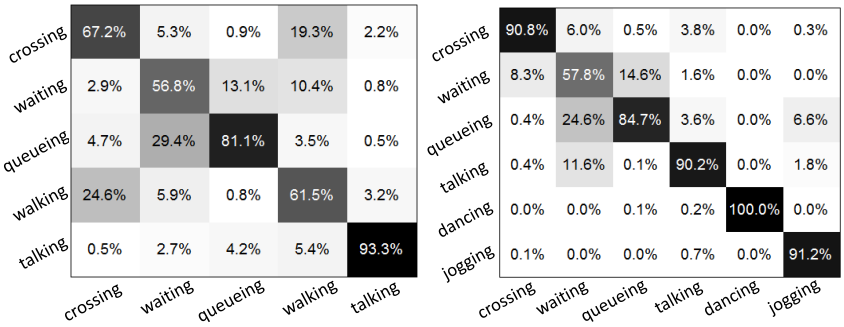

Fig. 3. Our confusion matrices for the 5-class [5] and the 6-class [7] datasets. The confusion matrices were obtained using the full model. Our classification accuracy is 72.0% on the 5-class dataset and 85.8% on the 6-class dataset.

Approach/Dataset	5 Activities	6 Activities
AC [1]	68.2%	-
STV+MC [5]	65.9%	-
RSTV [7]*	67.2%	71.7%
RSTV+MRF [7]*	70.9%	82.0%
Unary (AC) [8]	68.8%	81.5%
AC+Track Cues [8]	70.9%	83.7%
AC+Frame Cues	70.7%	84.8%
AC+Full Model	**72.0%**	**85.8%**

Table 1. A comparison of classification accuracies of the state-of-the-art methods on the two datasets. Our full model outperforms previous approaches and can be solved deterministically with some global optimality guarantees. * The approach of [7] employs additional trajectory information in training, including the 3D location and the pose of every person.

Acknowledgements

This research was partially supported by ONR MURI grant N000141010934 and by a grant from Siemens Corporate Research in Princeton, NJ.

References

1. Lan, T., Wang, Y., Mori, G., Robinovitch, S.N.: Retrieving actions in group contexts. In: SGA. (2010)
2. Marszalek, M., Laptev, I., Schmid, C.: Actions in context. In: CVPR. (2009)
3. Gupta, A., Davis, L.S.: Objects in action: An approach for combining action understanding and object perception. In: CVPR. (2007)
4. Yao, B., Fei-Fei, L.: Modeling mutual context of object and human pose in human-object interaction activities. CVPR (2010)

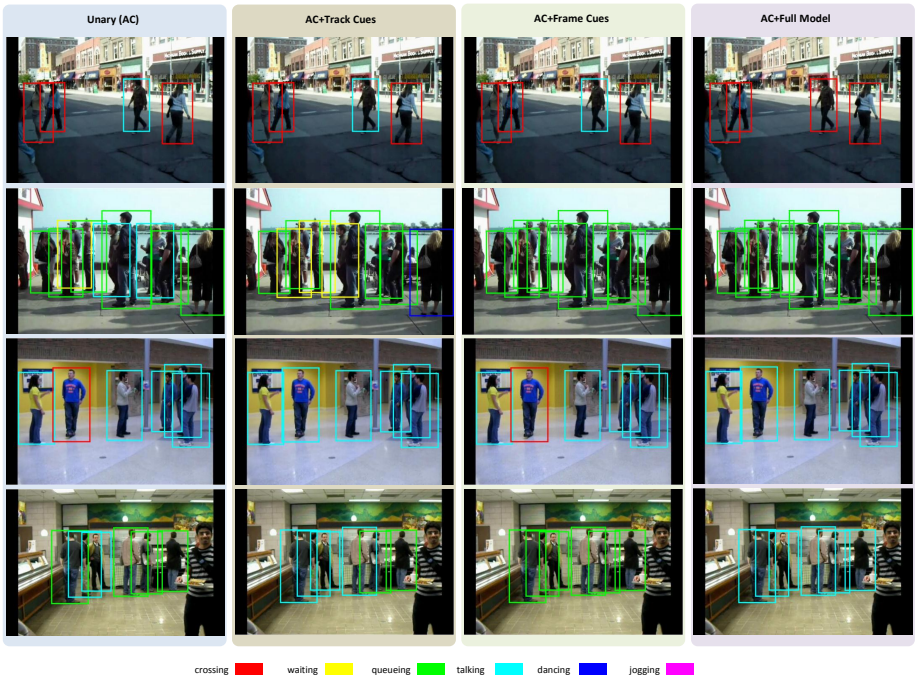

Fig. 4. Qualitative results of our model. The four columns represent the results using our unary potentials only, the track cues, the frame cues, and the full model respectively. The first row is a case where the full model, combining both cues, outperforms using either the frame cues or the track cues in isolation. In the second row the track cues degraded the results of the unary potentials due to identity matching inaccuracies in the busy scene, but the full model still yielded a perfect classification result. The frame cues were not able to fix classifier errors in the third row, but the full model leverged tracking and reported accurate results. Finally, the final row is a failure case where the full model reinforced the wrong result, classifying everyone incorrectly, even though the frame cues were successful.

5. Choi, W., Shahid, K., Savarese, S.: What are they doing?: Collective activity classification using spatio-temporal relationship among people. In: VS. (2009)
6. Lan, T., Wang, Y., Yang, W., Mori, G.: Beyond actions: Discriminative models for contextual group activities. In: NIPS. (2010)
7. Choi, W., Shahid, K., Savarese, S.: Learning context for collective activity recognition. In: CVPR. (2011)
8. Khamis, S., Morariu, V.I., Davis, L.S.: A flow model for joint action recognition and identity maintenance. In: CVPR. (2012)
9. Xiang, T., Gong, S.: Beyond tracking: modelling activity and understanding behaviour. IJCV **67** (2006) 21–51
10. Hakeem, A., Shah, M.: Learning, detection and representation of multi-agent events in videos. AI (2007)

11. Ryoo, M.S., Aggarwal, J.K.: Stochastic representation and recognition of high-level group activities. IJCV **93** (2010) 183–200
12. Gupta, A., Srinivasan, P., Shi, J., Davis, L.S.: Understanding videos, constructing plots learning a visually grounded storyline model from annotated videos. In: CVPR. (2009)
13. Morariu, V.I., Davis, L.S.: Multi-agent event recognition in structured scenarios. In: CVPR. (2011)
14. Brendel, W., Todorovic, S., Fern, A.: Probabilistic event logic for interval-based event recognition. In: CVPR. (2011)
15. Felzenszwalb, P., McAllester, D., Ramanan, D.: A discriminatively trained, multiscale, deformable part model. In: CVPR. (2008)
16. Zhang, L., Li, Y., Nevatia, R.: Global data association for multi-object tracking using network flows. In: CVPR. (2008)
17. Pirsiavash, H., Ramanan, D., Fowlkes, C.: Globally-optimal greedy algorithms for tracking a variable number of objects. In: CVPR. (2011)
18. Berclaz, J., Fleuret, F., Tretken, E., Fua, P.: Multiple object tracking using k-shortest paths optimization. PAMI **33** (2011) 1806–1819
19. Shitrit, H.B., Berclaz, J., Fleuret, F., Fua, P.: Tracking multiple people under global appearance constraints. In: ICCV. (2011)
20. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR. (2005)
21. Bertsekas, D.: Nonlinear Programming. Athena Scientific (1999)
22. Komodakis, N., Paragios, N., Tziritas, G.: Mrf optimization via dual decomposition: Message-passing revisited. In: ICCV. (2007)
23. Pearl, J.: Reverend bayes on inference engines: A distributed hierarchical approach. In: AAAI. (1982) 133–136
24. Gamarnik, D., Shah, D., Wei, Y.: Belief propagation for min-cost network flow: convergence & correctness. In: SODA. (2010)
25. Sutton, C., McCallum, A.: Piecewise training for undirected models. In: UAI. (2005)
26. Shotton, J., Winn, J., Rother, C., Criminisi, A.: Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In: ECCV. (2006)
27. Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: Liblinear: A library for large linear classification. JMLR **9** (2008) 1871–1874
28. Brendel, W., Amer, M., Todorovic, S.: Multiobject tracking as maximum-weight independent set. In: CVPR. (2011)
29. Weinberger, K.Q., Saul, L.K.: Fast solvers and efficient implementations for distance metric learning. In: ICML. (2008)
30. Gonfaus, J.M., Boix, X., de Weijer, J.V., Bagdanov, A.D., Serrat, J., Gonzàlez, J.: Harmony potentials for joint classification and segmentation. In: CVPR. (2010)
31. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: A local svm approach. In: ICPR. (2004)
32. Blank, M., Gorelick, L., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. In: ICCV. (2005)