Learning Latent Constituents for Recognition of Group Activities in Video

Borislav Antic and Björn Ommer

HCI & IWR, University of Heidelberg, Germany
borislav.antic@iwr.uni-heidelberg.de, ommer@uni-heidelberg.de

Abstract. The collective activity of a group of persons is more than a mere sum of individual person actions, since interactions and the context of the overall group behavior have crucial influence. Consequently, the current standard paradigm for group activity recognition is to model the spatiotemporal pattern of individual person bounding boxes and their interactions. Despite this trend towards increasingly global representations, activities are often defined by semi-local characteristics and their interrelation between different persons. For capturing the large visual variability with small semi-local parts, a large number of them are required, thus rendering manual annotation infeasible. To automatically learn activity constituents that are meaningful for the collective activity, we sample local parts and group related ones not merely based on visual similarity but based on the function they fulfill on a set of validation images. Then max-margin multiple instance learning is employed to jointly i) remove clutter from these groups and focus on only the relevant samples, ii) learn the activity constituents, and iii) train the multi-class activity classifier. Experiments on standard activity benchmark sets show the advantage of this joint procedure and demonstrate the benefit of functionally grouped latent activity constituents for group activity recognition.

Keywords: Group Activity Recognition, Latent Parts, Multiple-Instance Learning, Functional Grouping, Video Retrieval

1 Introduction

Over the last years there has been an ever growing interest in recognizing activities of groups of persons in video [24, 29, 32]. Whereas action recognition has a focus on the actions individual persons perform, group activities involve a group of people that perform the same or a related action in the scene such as talking to another. Thus, collective activity recognition is especially challenging as it depends on interactions and group behavior and so it is more than just the sum of individual person actions. For example the activity of lining up in a queue and the activity of waiting for a green light both exhibit the same individual standing actions and thus cannot be distinguished on that level. Conversely, analyzing the behavior of a single person benefits from recognition of group activities since noise and local occlusions can be overcome given the observations from the whole group. A main goal of this field is therefore to recognize the activity of a varying number of people and localize it on the level of person bounding boxes by identifying for each person what activity they partake in.

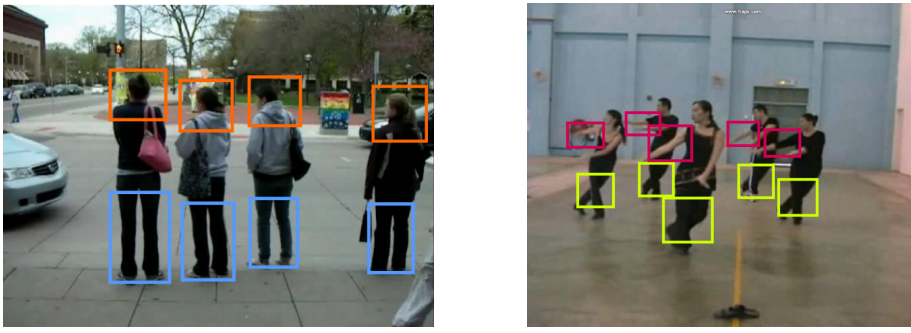


Fig. 1: Collective activities in videos can be recognized from semi-local characteristic parts that are grouped into a set of activity constituents by their common function on a validation set. Colors indicate different constituents and for legibility only a subset is shown.

The main theme is presently to estimate the activity of a complete person bounding box based on its visual features, motion, etc. [21, 8]. Thereafter contextual interactions are incorporated by re-classifying the activity at each bounding box based on the activities at neighboring boxes. The underlying rationale is that activities depend on the *overall* visual pattern of a bounding box and the activities in its neighborhood. Our hypothesis is that activities are much better characterized not on the level of bounding boxes but based on a large number of characteristic activity constituents and their interactions within the group. For example, if a detected constituent features a hand holding a tray, this indicates a person waiting in a queue.

Obviously, there are many different characteristic activity constituents and interactions thereof conceivable. Therefore, we cannot expect users to label them beforehand as in [5] and models limited to a small number of parts such as [14, 15] are not suited. Activities call for a large number of latent constituents with flexible but characteristic mutual arrangements and they need to be learned automatically without manual annotation. Training a large number of part classifiers, however, leaves only few samples per classifier. Moreover, many of these parts might be redundant (related regions coming from different persons) and recognition with too many redundant constituent classifiers is an unnecessary computational burden. Consequently, related parts need to be grouped into meaningful activity constituents before training a constituent classifier for each group. Fig. 1 shows a subset of the detected activity constituents for recognizing group activity in these two scenes. However, due to the curse of dimensionality, visual similarity becomes quickly unreliable in high-dimensional feature spaces. Consequently, the resulting groups would be cluttered and impair the subsequent learning of models for all parts within a group. As an example consider two patches on the legs of a person, one patch being larger and showing a shifted crop-out w.r.t. the other. Their feature vectors are vastly different, although both patches have the same function, i.e., they represent the gait of a walking person. Therefore, we match each patch to a set of validation images. Comparing their activations on a large number of validation samples, groups parts according to what *function* they have in these images, thus going beyond a mere visual

similarity. Since the grouped parts are sampled from noisy, cluttered scenes the resulting functional groups contain outliers. We thus frame the training of the constituent classifiers for each group as a multiple instance learning problem. Discriminating the activity constituent from negatives and removing irrelevant candidate instances from its group are tackled jointly. All of these activity constituents are then combined in a multi-class activity classifier, which is optimized together with its activity constituents to model their characteristic co-activations. Learning the activity constituents, selecting meaningful part samples for this learning, and optimizing the overall activity classifier are then coupled in a single objective function to represent concerted group activities.

2 Related Work

Action recognition in video has made significant progress over the last decade [29, 32]. A large body of literature on action recognition includes sparse feature representations by Dollar et al. [12], compositional models [27], action descriptions using correlations by Laptev et al. [22] or latent semantic analysis by Niebles et al. [26]. A very good summary on the action recognition techniques is provided in the survey [31].

Whereas action recognition is only concerned with actions performed by a single person or pairs of persons, recognition of group activities is about inferring the collective behavior of a group of persons that perform related actions. As the actions of individual persons are interrelated, the context of other persons in the scene is important for the overall activity. Pairwise interactions between persons in a scene are used for recognizing human activities in [28]. [3] propose a joint parsing of entire scenes for detecting abnormalities in video. Xiang and Gong [33] reason about temporal and causal correlations among scene events in a video. The recognition of group activities is very challenging because of the complex interactions between persons in a scene. Some authors proposed various contextual descriptors for representing the group activities [9, 19, 10]. Hierarchical models that unify individual actions, pairwise interactions and group activities were proposed in [20, 8].

Recent work on collective activity recognition aims to jointly solve the problems of group activity recognition and person tracking [18, 17, 7]. Choi and Savarese [7] leverage target interactions for guiding target associations, and use different levels of granularity for encoding the activities. Amer and Todorovic [1] allow for arbitrarily large numbers of activity participants and localize parts of the activity using a chain model of group activity. In [2] the authors use a three-layered AND-OR graph to jointly model group activities, individual actions, and participating objects. Khamis et al. [18] combine per-track and per-frame cues for action recognition and bounding box tracking.

Human body patches with similar pose and appearance have been obtained in Poselets [23] by a very detailed manual annotation of the joints of persons. In contrast to this tedious labeling task, our method automatically discovers functionally related parts and learns the latent constituents jointly with the activity classifier using multiple-instance learning. Related work on mid-level discriminative patches [30] and intermediate compositional parts [13] learn the patches independently from the overall classifier. Consequently, mid-level patches do not directly maximize the overall classification performance. In contrast, our approach jointly optimizes constituent parts and the group

Fig. 2: Reconstructing query image (left) using groups of parts found by visual similarity based clustering (middle) and using functional grouping (right). See text for the details.

activity classifier. Moreover, a functional grouping based on similarity of part activation patterns on a validation set leads to more consistent initial part clusters than those based on visual part similarity.

3 Learning Functional Constituents of Group Activities

Group activity recognition requires for each person to infer which collective activity they partake in. That is, for each person bounding box we seek the activity that is consistent with other interacting persons in the scene. To capture the peculiarities of activities we have to go beyond a mere representation on the level of bounding boxes and grasp the semi-local constituents of activities. Due to the large within-class variability this requires a large number of parts, so learning a separate model for each part is neither statistically feasible (few training samples in high-dimensional feature space) nor computationally practical (high complexity).

Our approach adaptively seeks the feasible middle-ground between two extremes: (i) a single or only few complex classifiers that try to capture all characteristics of activities with all their multi-modal variabilities, and (ii) an impossibly large number of separate classifiers for each part with too few training samples for each and redundancies between them. Thus we aim at grouping related semi-local parts from different training images according to the function they take in a set of validation images, so we can train a single, more powerful classifier for each group of related activity constituents. Functional grouping yields a candidate set of part instances per activity constituent, but groups may still contain clutter and spurious part samples. That is why classifiers for activities and their latent constituents are learned jointly by employing multiple-instance learning to select only the relevant instances for each constituent from its candidate set (see Fig. 3).

3.1 Functional Grouping of Part Instances

Our goal is to learn the latent constituents that make up the group activities without having manual annotations for these components. Constituents are semi-local regions that cover characteristic parts of persons and their surroundings. Due to their number and flexible arrangement they successfully bridge the gap between local features and the group activity with its large spatial extend. To this end we need to resolve the crucial limitations of present activity recognition systems: they are limited to only a small set

of predefined parts and require manually annotated training samples for each [8, 17, 21]. Therefore, we seek to automatically learn a large number of constituents that effectively capture the characteristics of an activity.

To learn the latent constituents of group activities, we first randomly sample a large number of parts R_i at different locations $\mathbf{x}_i \in \mathbb{R}^2$ and sizes $s_i \in \mathbb{R}_+$ within person bounding boxes so that there is a good coverage of all training instances of an activity class. To be comparable to previous work, we follow the same person detection approach as in [18, 21] to obtain the person boxes.

The probability of sampling a part R_i is inversely proportional to its overlap with already previously sampled parts $R_{i_1}, R_{i_2}, \ldots, R_{i_n}$,

$$\omega_i \propto \left(\max_n \frac{|R_i \cap R_{i_n}|}{|R_i \cup R_{i_n}|} + \epsilon \right)^{-1}. \tag{1}$$

Parts that are sampled from different bounding boxes do not have any pixel in common, so their intersection is zero. Therefore, regions that have not yet been sampled will have a high likelihood of being selected, thus improving the overall coverage.

A part instance R_i that is extracted from a person bounding box is described by the feature vector $\mathbf{f}_i \in \mathcal{F}$. To compare to previous work we utilize the same feature space $\mathcal{F}$ as in [7], a bag-of-feature (BoF) [12] and histograms of oriented gradients (HoG) features.

Despite variations in their feature vectors $\mathbf{f}_i$, many of the sampled parts are related instances with the same function or meaning for the overall activity. They are instances of the same component of an activity such as different images of a knee bent at right angle represent the same aspect of the *jogging* activity. Training a separate classifier for each sampled part instance is not feasible, since this would yield a large number of mutually redundant classifiers with few training data for each. Moreover, recognition would not be practicable, because a large set of related classifiers would all need to be evaluated. Thus we need to group functionally related parts that represent the same characteristic within the overall activity.

Inferring useful groups of parts is a challenging problem. Parts with the same meaning are commonly not close in the high-dimensional feature space of their descriptors. Consequently, feature similarities are quickly becoming arbitrary given only little noise or visual variation. The problem of visual grouping can, however, be circumvented by observing the activation pattern of each part on an independent validation set. Part instances that are simultaneously active or inactive on many validation samples in similar regions are related to the same target concept. These parts have the same function in explaining the group activity and so we refer to this process as *functional grouping* in contrast to a grouping based on feature descriptor similarity.

We first divide the positive training samples into a training and validation set T and V ($3/4$ vs. $1/4$). Then we find for each part $\{\mathbf{f}_i\}_{i \in T}$ from the training set its k best matches from the validation set $\{\mathbf{f}'_i\}_{i' \in V}$. To find these matching parts we employ an approximate nearest neighbor search [25]. Let $V_i \subseteq V$ refer to the k best matches for the i-th training part. The activation pattern of this training part is then given as a Gibbs

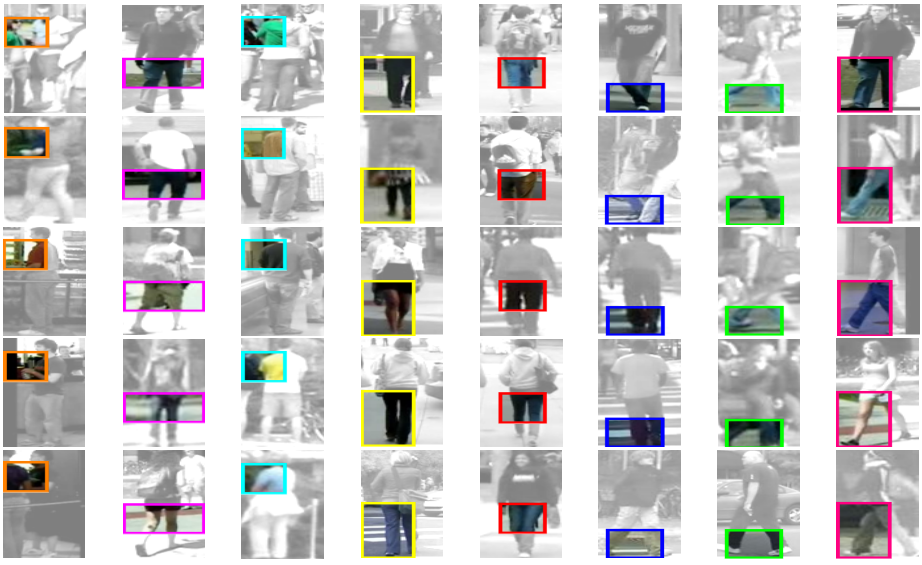


Fig. 3: Visualization of latent constituents of the group activities from the 5-class datasets. Columns show eight randomly drawn activity constituents. For each constituent five of its randomly sampled parts are shown.

distribution over the validation set, with support V_i,

$$\alpha_i(v) := \begin{cases} \frac{1}{Z_i} \exp\big(-\beta \Delta(\mathbf{f}_i, \mathbf{f}_v)\big), & \text{if } v \in V_i \\ 0, & \text{if } v \in V \setminus V_i \end{cases} \tag{2}$$

where $\Delta(\cdot,\cdot)$ is a distance function between the sampled parts which combines the distance $d(\cdot,\cdot)$ in the feature space $\mathcal{F}$ with the differences in parts' location and size,

$$\Delta(\mathbf{f}_i, \mathbf{f}_v) := d(\mathbf{f}_i, \mathbf{f}_v) + \lambda_x \|\mathbf{x}_i - \mathbf{x}_v\| + \lambda_s |s_i - s_v|. \tag{3}$$

To balance individual distances, we use λ_x and λ_s as the quotients of the feature variance to the variances of parts' location and size, respectively. Z_i is the partition function of the Gibbs distribution,

$$Z_i = \sum_{v \in V_i} \exp\Big(-\beta \Delta(\mathbf{f}_i, \mathbf{f}_v)\Big). \tag{4}$$

To group related parts into a set of meaningful activity constituents, we perform agglomerative clustering (Ward's Method) based on the functional relatedness of parts $\mathbf{f}_i$ and $\mathbf{f}_{i'}$, $i, i' \in T$, estimated by calculating the distance between part activation patterns α_i and $\alpha_{i'}$. As a result of clustering, training instances of all parts $\{R_i\}_{i \in T}$ are divided into disjoint groups of related parts $\{T_1, T_2, \ldots, T_G\}$. Each group T_g contains part instances that are related to the same concept. Now we can train a discriminative classifier with weight vector $\mathbf{w}_g$ for g-th activity constituent that separates its positive

training samples in T_g from a set of random negatives N_g drawn from other activities. Note that these classifiers do not distinguish between activities but only detect the presence of certain characteristic components of an activity.

Activity Reconstruction with Functional Groups: Let us now reconstruct a query image using the functional groups, Fig. 2. For each part sampled in the query image we infer the activity constituent g with maximal score from its linear classifier $\mathbf{w}_g$ on that part. Then we randomly select a training part $i \in T_g$ from its functional group. The final reconstructed image is then obtained as the weighted average (weighted with the classifier scores) of all individual sampled parts i which are then placed at locations of the original parts of the query image. As an additional experiment we cluster parts according to their mere feature similarity (also using Ward's method) and sample from the resulting groups. As illustrated in Fig. 2 the reconstruction from functional groups captures the characteristics of activities, whereas the reconstruction based on the visual similarity leads to fuzzy clusters and therefore looses important details during reconstruction.

3.2 From Constituent Classifiers to Activity Classification

Since our final goal is not only to discover group activities, but also to localize them by labeling each person with the activity it is part of, for each person the context of other persons in the group matters. A commonly used representation for action context is the descriptor from [19]. Here we extend this representation and employ it not once per person as in [21, 17], but for each activity constituent g. Consequently, we are establishing context on the level of constituents rather than merely between person bounding boxes.

The action context descriptor for a constituent g divides the spatio-temporal volume around the j-th person detection that this constituent belongs to into disjoint subvolumes $\{\mathcal{N}_{j,1}, \ldots, \mathcal{N}_{j,M}\}$ (parameters as in [19]). Then the constituent classifier $\mathbf{w}_g$ searches for other occurrences of the same constituent in each of the neighboring regions $\mathcal{N}_{j,m}$. The response for $\mathcal{N}_{j,m}$ is the maximal score of the classifier on all $\mathbf{f}_i$ within $\mathcal{N}_{j,m}$,

$$q_j(m, g) = \max_{i \in \mathcal{N}_{j,m}} \mathbf{w}_g^\top \mathbf{f}_i. \tag{5}$$

q then expresses relationships between different detections of a constituent in the neighborhood.

The joint activity representation $\mathbf{q}_j$ for person j is obtained by concatenating score function values $q_j(m, g)$ for all sub-volumes m and constituents g. The activity score of a person j for group activity $a \in \mathcal{A}$ is calculated using a linear classifier $\mathbf{w}_a^\top \mathbf{q}_j$, where $\mathcal{A}$ is the set of all group activity labels and $\mathbf{w}_a$ is a hyperplane that separates instances with activity label a from instances of other group activities, i.e., this is a multi-class classification problem which will be discussed in Sect. 3.4. Due to the max operation in Eq. 5, the group activity is more than just a mere sum of activity constituents.

3.3 Inference in Novel Query Scenes

In the recognition phase, we need to detect group activities $a \in \mathcal{A}$ and localize them on the level of bounding boxes. To get this process started, person bounding boxes are first

	5 Activities	6 Activities
Baseline (no latent constituents)	70.4%	83.3%
Constituents from Visual Grouping	71.7%	87.4%
Lat. Constituents + Functional Grouping	74.1%	89.2%
Lat. Constit. + Func. Grouping + Dense Traject.	75.1%	90.1%

Table 1: Comparing the latent constituent model with the baseline method (no constituents, parts are whole bounding boxes), and method that uses constituents learned by visual grouping. The gain of the latent constituent approach with functional grouping over the baseline are 3.7% and 5.9% on 5-class and 6-class benchmark sets, respectively. Last row shows the results of the classification if standard feature representation for the latent constituents is augmented with dense trajectories [32].

localized as in [18, 21]. Then the constituent classifiers $\mathbf{w}_g$ detect occurrences and then compute the context scores $q_j(m, g)$ for each person j using Eq. 5. The group activity which a person j belongs to results from applying the overall activity classifiers,

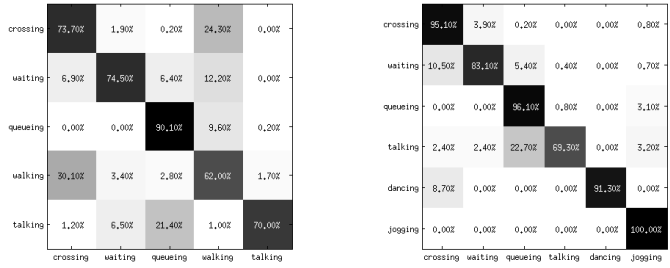
$$a_j = \underset{a' \in \mathcal{A}}{\operatorname{argmax}} \, \mathbf{w}_{a'}^{\top} \mathbf{q}_j. \tag{6}$$

3.4 Joint Learning of Group Activity and Constituent Classifiers

Let us now jointly learn the weights $\mathbf{w}_g$ of all constituent classifiers and of the overall activity models $\mathbf{w}_a$ for all activity classes by adopting a max-margin rationale. The $\mathbf{w}_g$ should discriminate positive instances $\{\mathbf{f}_i\}_{i \in T_g}$ of a constituent group from negative ones $\{\mathbf{f}_i\}_{i \in N_g}$. However, since the training set T contains clutter (regions that are noisy, contain outliers, or have been sampled from uninformative areas), the decomposition of T into disjoint sets T_g by means of functional grouping in Sect. 3.1 still contains these outliers. Identifying these outliers with the constituent classifiers $\mathbf{w}_g$ and training the $\mathbf{w}_g$ with only the remaining meaningful instances are interrelated problems that need to be solved jointly. We follow a Multiple Instance Learning (MIL) approach [16, 4], which selects positive instances from a positive bag that are used for training a discriminative classifier. Our approach is motivated by the AL-SVM method [16] that uses deterministic annealing to train the classifier and find the unknown instance labels with the entropy regularizer $\mathcal{H}(\cdot)$. We associate a probability $p_i \in [0, 1]$ to each part $i \in T_g$, that indicates how meaningful i is for learning the g-th constituent classifier, i.e., zero implies an outlier. The MIL objective is then to find at least $\rho|T_g|$ positive examples in a functional group g (we simply set $\rho = 0.7$ and observed only little influence when changing ρ), so that the hinge loss error $\ell(\mathbf{w}_g^{\top}\mathbf{f}_i) = \max(0, 1 - \mathbf{w}_g^{\top}\mathbf{f}_i)$ of a regularized classifier $\mathbf{w}_g$ is minimal,

$$\min_{\mathbf{w}_g, p_i} \|\mathbf{w}_g\|^2 + C_g\Big\{\sum_{i \in T_g} p_i \ell(\mathbf{w}_g^{\top}\mathbf{f}_i) + \sum_{i \in N_g} \ell(-\mathbf{w}_g^{\top}\mathbf{f}_i)\Big\} - T \sum_{i \in T_g} \mathcal{H}(p_i), \tag{7}$$

$$\text{s.t.} \sum_{i \in T_g} p_i \geq \rho|T_g| \quad \wedge \quad p_i \in [0, 1], \forall i. \tag{8}$$



(a) (b)

Fig. 4: a) Per-class categorization results (confusion matrix) on the 5-class collective activity dataset [9]. The results are obtained by the proposed latent constituent method with functional grouping using standard features. The average classification accuracy is 74.1%. b) The confusion matrix for the 6-class benchmark set [10]. Average classification accuracy is 89.2%

Activity classification with $\mathbf{w}_a$ (Eq. 6) now depends on all $\mathbf{w}_g$ and their context scores $\mathbf{q}_j$ (Eq. 5). Thus we should not only optimize the $\mathbf{w}_a$, but jointly estimate the $\mathbf{w}_g$, so that they improve the discrimination between activities in Eq. 6. Following the standard training protocol of [9], training bounding boxes are provided for each person j, together with their activity labels $y_{j,a} \in \{1, -1\}$, where $y_{j,a} = 1$ if person j participates in activity a. Thus, optimizing $\mathbf{w}_g$ and finding meaningful representative parts $i \in T_g$ with MIL in Eq. 7 and max-margin training of $\mathbf{w}_a$ are coupled,

$$\min_{\mathbf{w}_a, \mathbf{w}_g, p_i} \sum_a \Big\{ \|\mathbf{w}_a\|^2 + C_a \sum_j \ell(y_{j,a} \mathbf{w}_a^\top \mathbf{q}_j) \Big\} + \sum_g \Big\{ \|\mathbf{w}_g\|^2 + C_g \big[\sum_{i \in T_g} p_i \ell(\mathbf{w}_g^\top \mathbf{f}_i) + \sum_{i \in N_g} \ell(-\mathbf{w}_g^\top \mathbf{f}_i) \big] - T \sum_{i \in T_g} \mathcal{H}(p_i) \Big\}, \tag{9}$$

$$\text{s.t.} \sum_{i \in T_g} p_i \geq \rho |T_g|, \forall g \quad \wedge \quad p_i \in [0, 1], \forall i. \tag{10}$$

The joint optimization problem in Eq. 9 is non-convex, therefore we solve it using alternating optimization.

(i) To find the most meaningful samples in each group T_g we need to solve for the probabilities $p_i, i \in T_g$. Finding the optimal value of the Lagrangian,

$$\min_{\{p_i\}_{i \in T_g}} \mathcal{L}(\{p_i\}_{i \in T_g}, \lambda) = C_g \sum_{i \in T_g} p_i \ell(\mathbf{w}_g^\top \mathbf{f}_i) + T \sum_{i \in T_g} \big(p_i \log p_i + (1 - p_i) \log(1 - p_i) \big) - \lambda \big(\sum_{i \in T_g} p_i - \rho |T_g| \big), \text{ s.t. } \lambda \geq 0, \tag{11}$$

	5 Activities	6 Activities
AC [19]	68.2%	-
STV + MC [9]	65.9%	-
RSTV [10]	67.2%	71.2%
RSTV + MRF [10]	70.9%	82.0%
AC (Unary) [18]	68.8%	81.5%
AC + Track Cues [18]	70.9%	83.7%
AC + Frame + Track Cues [17]	72.0%	85.8%
Unified Track. + Recognit. [7]	74.4%	-
Latent Constituents	74.1%	89.2%
Latent Constituents + Dense Traject.	**75.1%**	**90.1%**

Table 2: Comparison of the state-of-the-art methods for group activity recognition on 5-class and 6-class datasets [9, 10]. Our latent constituents achieve best performance of 89.2% on the 6-class dataset (a gain of 3.4% over state-of-the-art [17]), and its performance on 5-class dataset has a comparable performance of 74.1% to the state-of-the-art [7] using standard features and performance increases to 90.1% and 75.1% respectively with dense trajectories [32].

together with the Karush-Kuhn-Tucker conditions following from Eq. 10, the solution can be derived in analytical form

$$p_i = \sigma\Big(-\frac{C_g \ell(\mathbf{w}_g^\top \mathbf{f}_i)}{T}\Big) \cdot \max\Big(\rho|T_g|\Big\{\sum_{i\in T_g} \sigma\Big(-\frac{C_g \ell(\mathbf{w}_g^\top \mathbf{f}_i)}{T}\Big)\Big\}^{-1}, 1\Big), \forall i \in T_g, \quad (12)$$

where $\sigma(x) = \big(1+\exp(-x)\big)^{-1}$ is the sigmoid function.

(ii) Now we can discuss the training of constituent classifiers $\mathbf{w}_g$. Note from Eq. 5 that the score function related to person j can be written in linear form $q_j(m,g) = \mathbf{w}_g^\top \mathbf{f}_{i_g^*}$, where $i_g^* = \operatorname{argmax}_{i\in\mathcal{N}_{j,m}} \mathbf{w}_g^\top \mathbf{f}_i$. By concatenating the features $\mathbf{f}_{i_g^*}$ over all neighbors m into a matrix $\mathbf{F}_j$ the score vector becomes $\mathbf{q}_j = \mathbf{F}_j^\top \mathbf{w}_g$. The constituent classifier is solved as a convex optimization problem using ILOG CPLEX solver for the problem

$$\min_{\mathbf{w}_g} \|\mathbf{w}_g\|^2 + C_g\big[\sum_{i\in T_g} p_i \ell(\mathbf{w}_g^\top \mathbf{f}_i) + \sum_{i\in N_g} \ell(-\mathbf{w}_g^\top \mathbf{f}_i)\big] + C_a \sum_j \ell\big(y_{j,a} \mathbf{w}_g^\top (\mathbf{F}_j \mathbf{w}_a)\big). \quad (13)$$

(iii) Based on p_i and $\mathbf{w}_g$, optimizing $\mathbf{w}_a$ becomes a multi-class linear SVM problem of the same formulation as [6], which is solved using LIBLINEAR,

$$\min_{\mathbf{w}_a} \sum_a \|\mathbf{w}_a\|^2 + C_a \sum_j \ell(y_{j,a} \mathbf{w}_a^\top \mathbf{q}_j). \quad (14)$$

We alternate between these three steps until convergence that is typically achieved within on the order of ten iterations. Visualization of the learned latent constituents for a benchmark set that we use in our experiments is given in Fig. 3.

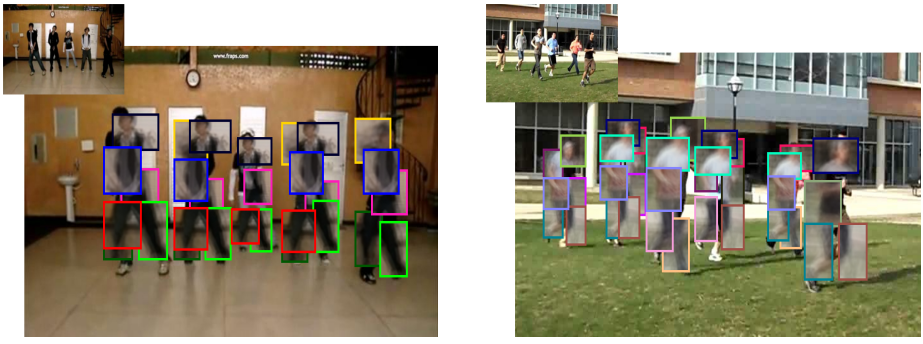


Fig. 5: Visualization of detected latent constituents of two example group activities. Detected parts that correspond to the same latent constituent are framed in the same color. Each part is then visualized by averaging over the training patches that define the latent constituent.

4 Experimental Results

4.1 Experimental Protocol

We evaluate our approach on two standard benchmark sets for group activity recognition that were recently proposed in [9, 10]. Both datasets are recorded with a hand-held camera in realistic indoor and outdoor environments. Popular action recognition datasets such as KTH, Weizmann, Hollywood or UCF Sports are not appropriate for evaluation, since they contain only isolated human actions, but not any group activity.

The first benchmark set [9] consists of 44 videos showing 5 group activities (*crossing*, *standing*, *queueing*, *walking*, and *talking*). The length of videos ranges from less than 100 frames to more than 2000 frames. The second benchmark set [10] contains 72 videos and involves 6 group activity classes. This dataset is created by augmenting the first dataset, adding *dancing* and *jogging* and removing *walking* categories. We follow the common experimental protocol [21, 17, 18, 7] that provides bounding boxes for persons in the training set with corresponding group activity labels and uses a leave-one-out framework for testing. To assess localized group activity recognition performance, we follow the standard protocol and evaluate on a per-bounding-box level.

Following recent practice in the group activity recognition literature [7, 8], we use the combination of HOG [11] and bag-of-feature (BoF) [12] features to represent the constituents of group activities, c.f. Sect. 3.1. We also adhere to a common practice in group activity recognition [7, 17] that associates object detections from different frames of a video by object tracking. Learning begins by randomly sampling (Eq. 1) 10000 parts from training person bounding boxes. The functional grouping of Sect. 3.1 then creates $G = 100$ constituents. Max-margin MIL training learns the importance $\mathbf{w}_a$ of each constituent for group activity recognition. Constituents with small value in the activity classifier $\mathbf{w}_a$ can be skipped during inference. The number of constituents that is used for testing is chosen so that 80% of the classifier's energy $\|\mathbf{w}_a\|^2$ is retained.

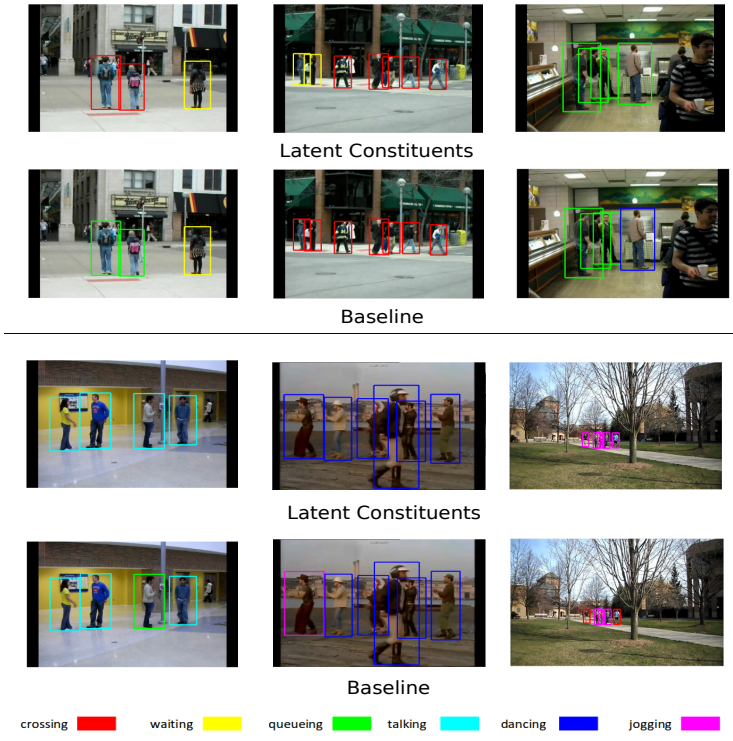


Fig. 6: Visualization of the classification results on test videos comparing our latent constituent model and the baseline approach that uses bounding boxes as is common but no further constituents. Frames are taken from the test videos.

Typically around 50 constituents are retained after such selection process, which is why we observed only little influence when changing G.

4.2 Group Activity Recognition

Fig. 5 shows latent constituents for two example group activities (*dancing* and *jogging*). Different colors represent different latent constituents. Moreover, each constituent is visualized by averaging the training patches that defined it during learning.

Table 1 presents the results of group activity recognition on the benchmark sets. We compare our latent constituent approach to the baseline method that does not use any latent constituents, but is merely based on the whole bounding box of a person as in other activity approaches such as [18, 19]. Our method achieves a gain of 3.7% over the baseline for the 5-class dataset, and 5.9% for the 6-class dataset. We also compare our latent constituents found by functional grouping and MIL with the baseline approach of performing visual clustering, i.e., directly clustering the part feature vectors. Our method achieves an improvement of 2.4% over the visual grouping method for the 5-

Fig. 7: Sampling parts and reconstructing (explaining) them from the training parts that belong to the same constituents as the original part. Samples are taken from *waiting*, *crossing* and *walking* activities.

class dataset, and 1.8% on the 6-class dataset. We also conduct an experiment in which we update the standard feature representation used for the latent constituents with recent dense trajectory features [32]. We obtain an increase in the performance and achieve 75.1% for the 5-class dataset and 90.1% for the 6-class dataset.

Per-category classification results of our latent constituent method based on standard features on the two benchmark sets are given in confusion matrices of Fig. 4a and 4b. We notice that the greatest confusion is between *walking* and *crossing* activities, because they have many constituents in common. *Talking* activity is also often misclassified as *queueing* because of their similarity.

We next compare our performance with other state-of-the-art results in Table 2. Our method achieves an average accuracy of 74.1% on the 5-class dataset, that is almost the same as the best performing method [7] that uses additional manual labels. Our method achieves 89.2% on the 6-class dataset, which is 3.4% better than the state-of-the-art. The visual results of our group activity classification and of the baseline method are shown in Fig. 6.

Activity Reconstruction with Latent Constituents: Fig. 2 shows reconstructions provided by groups of parts obtained by functional grouping with those coming from groups of parts obtained by clustering with visual similarity, i.e., clustering the part appearance features. Fig. 7 presents additional reconstructions for person bounding boxes in several activity classes: *crossing*, *waiting* and *walking*. Each image part is replaced by a randomly sampled part from the training set that belongs to the same constituent as the original image part. The final reconstructed image is obtained by averaging. Again one can see that key characteristics of an activity class are captured by latent constituents.

5 Conclusion

This paper has demonstrated that activity recognition significantly benefits from modeling human behavior using a large number of semi-local parts and their interaction between persons. Learning the underlying classifier becomes feasible by grouping functionally related parts into activity constituents and removing clutter with multiple instance learning while training constituent classifiers and the multi-class activity model. The approach has shown a significant performance gain on standard activity benchmarks.

Acknowledgements

This work has been partially supported by the Ministry for Science, Baden-Wuerttemberg and German Research Foundation (DFG) within the program Spatio-/Temporal Graphical Models and Applications in Image Analysis, grant GRK 1653. The authors would also like to thank Timo Milbich for additional support and Till Kroeger for helpful discussions.

References

1. Amer, M.R., Todorovic, S.: A chains model for localizing participants of group activities in videos. In: ICCV. pp. 786–793 (2011)
2. Amer, M.R., Xie, D., Zhao, M., Todorovic, S., Zhu, S.C.: Cost-sensitive top-down/bottom-up inference for multiscale activity recognition. In: ECCV (4). pp. 187–200 (2012)
3. Antic, B., Ommer, B.: Video parsing for abnormality detection. In: ICCV. pp. 2415–2422 (2011)
4. Antic, B., Ommer, B.: Robust multiple-instance learning with superbags. In: ACCV, Lecture Notes in Computer Science, vol. 7725, pp. 242–255. Springer Berlin Heidelberg (2012)
5. Bourdev, L., Maji, S., Brox, T., Malik, J.: Detecting people using mutually consistent poselet activations. In: European Conference on Computer Vision (ECCV) (2010)
6. Chapelle, O., Keerthi, S.S.: Multi-class feature selection with support vector machines. In: Proc. of the American Statistical Assoc. (2008)
7. Choi, W., Savarese, S.: A unified framework for multi-target tracking and collective activity recognition. In: ECCV (2012)
8. Choi, W., Savarese, S.: Understanding collective activities of people from videos. Pattern Analysis and Machine Intelligence (99), 1–1 (2013)
9. Choi, W., Shahid, K., Savarese, S.: What are they doing? : Collective activity classification using spatio-temporal relationship among people. In: Proc. of 9th International Workshop on Visual Surveillance (VSWS09) in conjuction with ICCV (2009)
10. Choi, W., Shahid, K., Savarese, S.: Learning context for collective activity recognition. In: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (2011)
11. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Schmid, C., Soatto, S., Tomasi, C. (eds.) International Conference on Computer Vision & Pattern Recognition. vol. 2, pp. 886–893 (2005)
12. Dollár, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features. In: Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on. pp. 65–72. IEEE (2005)
13. Eigenstetter, A., Takami, M., Ommer, B.: Randomized Max-Margin Compositions for Visual Recognition. In: CVPR - International Conference on Computer Vision and Pattern Recognition. Columbus, USA (2014)
14. Felzenszwalb, P.F., Girshick, R.B., McAllester, D.A., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE Trans. Pattern Anal. Mach. Intell. 32(9), 1627–1645 (2010)
15. Fergus, R., Perona, P., Zisserman, A.: Object class recognition by unsupervised scale-invariant learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. vol. 2, pp. 264–271 (2003)
16. Gehler, P.V., Chapelle, O.: Deterministic annealing for multiple-instance learning. In: International conference on artificial intelligence and statistics. pp. 123–130 (2007)

17. Khamis, S., Morariu, V.I., Davis, L.S.: Combining per-frame and per-track cues for multi-person action recognition. In: European Conference on Computer Vision (2012)
18. Khamis, S., Morariu, V.I., Davis, L.S.: A flow model for joint action recognition and identity maintenance. In: IEEE Conference on Computer Vision and Pattern Recognition (2012)
19. Lan, T., Wang, Y., Mori, G., Robinovitch, S.: Retrieving actions in group contexts. In: International Workshop on Sign Gesture Activity (2010)
20. Lan, T., Wang, Y., Yang, W., Mori, G.: Beyond actions: Discriminative models for contextual group activities. In: Advances in Neural Information Processing Systems (NIPS) (2010)
21. Lan, T., Wang, Y., Yang, W., Robinovitch, S., Mori, G.: Discriminative latent models for recognizing contextual group activities. IEEE Transactions on Pattern Analysis and Machine Intelligence (2012)
22. Laptev, I., Marszałek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: Conference on Computer Vision & Pattern Recognition (2008)
23. Maji, S., Bourdev, L., Malik, J.: Action recognition from a distributed representation of pose and appearance. In: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) (2011)
24. Marszałek, M., Laptev, I., Schmid, C.: Actions in context. In: Conference on Computer Vision & Pattern Recognition (2009)
25. Muja, M., Lowe, D.G.: Fast approximate nearest neighbors with automatic algorithm configuration. In: International Conference on Computer Vision Theory and Application (VISSAPP'09). pp. 331–340. INSTICC Press (2009)
26. Niebles, J.C., Wang, H., Fei-Fei, L.: Unsupervised learning of human action categories using spatial-temporal words. Int. J. Comput. Vision 79(3), 299–318 (2008)
27. Ommer, B., Mader, T., Buhmann, J.M.: Seeing the objects behind the dots: Recognition in videos from a moving camera. International Journal of Computer Vision 83(1), 57–71 (2009)
28. Ryoo, M.S., Aggarwal, J.K.: Stochastic representation and recognition of high-level group activities. International Journal of Computer Vision 93(2), 183–200 (2011)
29. Schüldt, C., Laptev, I., Caputo, B.: Recognizing human actions: A local svm approach. In: ICPR (3). pp. 32–36 (2004)
30. Singh, S., Gupta, A., Efros, A.A.: Unsupervised discovery of mid-level discriminative patches. In: European Conference on Computer Vision (2012)
31. Turaga, P.K., Chellappa, R., Subrahmanian, V.S., Udrea, O.: Machine recognition of human activities: A survey. IEEE Trans. Circuits Syst. Video Techn. 18(11), 1473–1488 (2008)
32. Wang, H., Kläser, A., Schmid, C., Liu, C.L.: Action Recognition by Dense Trajectories. In: IEEE Conference on Computer Vision & Pattern Recognition. pp. 3169–3176. Colorado Springs, United States (2011)
33. Xiang, T., Gong, S.: Beyond tracking: Modelling activity and understanding behaviour. International Journal of Computer Vision 67(1), 21–51 (2006)