A Fully Connected Model for Consistent Collective Activity Recognition in Videos

Takuhiro Kaneko*, Masamichi Shimosaka, Shigeyuki Odashima, Rui Fukui, Tomomasa Sato

Department of Mechano-Informatics, the University of Tokyo, Tokyo, Japan

Abstract

We propose a novel method for consistent collective activity recognition in video images. Collective activities are activities performed by multiple persons, such as queuing in a line, talking together, and waiting at an intersection. Since it is often difficult to differentiate between these activities using the appearance of only an individual person, the models proposed in recent studies exploit the contextual information of other people nearby. However, these models do not sufficiently consider the spatial and temporal consistency in a group (e.g., they consider the consistency in only the adjacent area), and therefore, they cannot effectively deal with temporary misclassification or simultaneously consider multiple collective activities in a scene. To overcome this drawback, this paper describes a method to integrate the individual recognition results via fully connected conditional random fields (CRFs), which consider all the interactions among the people in a video clip and alter the interaction strength in accordance with the degree of their similarity. Unlike previous methods that restrict the interactions among the

*Corresponding author
Email address: kaneko@ics.t.u-tokyo.ac.jp (Takuhiro Kaneko)

people heuristically (e.g., within a constant area), our method describes the "multi-scale" interactions in various features, i.e., position, size, motion, and time sequence, in order to allow various types, sizes, and shapes of groups to be treated. Experimental results on two challenging video datasets indicate that our model outperforms not only other graph topologies but also state-of-the art models.
Keywords:
Collective activity recognition, Fully connected model, CRFs, Spatial and temporal consistency

---

1. Introduction

Vision-based human activity recognition is of scientific and practical importance, and has been actively studied in the research field of computer vision. Many previous studies focused on recognizing actions performed by a single person in a video clip (Blank et al., 2005; Niebles et al., 2006; Schuldt et al., 2004). However, in real-world applications, such as surveillance monitoring, the previous methods are inapplicable, since human actions are rarely performed by a single person, but instead by multiple persons. For example, it is difficult to differentiate between the activities of the two persons shown in Fig. 1(a), by considering the appearance of the individual person. In order to recognize activities performed by multiple persons, which we call "collective activities," it is necessary to exploit the contextual information of the people nearby. When we have identified the activities of people nearby, it immediately becomes clear that the left person in Fig. 1(a) is queuing and the right person is talking, as shown in Fig. 1(b).

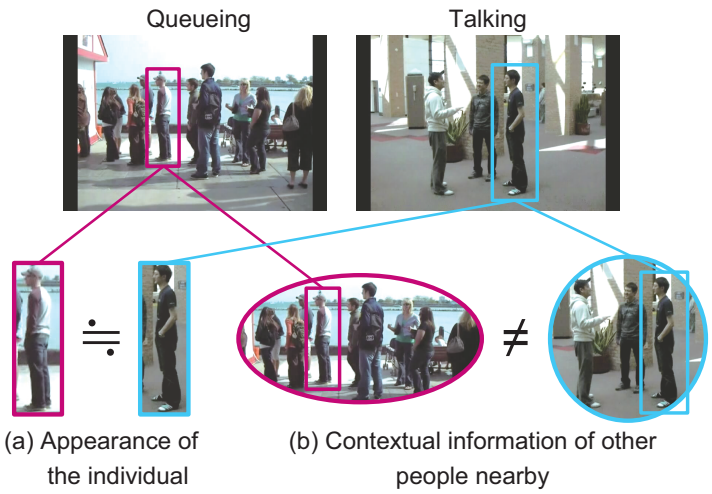

Figure 1: Useful contexts for collective activity recognition. It is often difficult to differentiate between collective activities by the appearance of only an individual person (a). When we have identified the activities of people nearby, it immediately becomes clear that the left person is queuing and the right person is talking (b).

In some recent studies, methods have been proposed for collective activity recognition using the contextual information of people nearby. Choi et al. (2009), Lan et al. (2010a), and Kaneko et al. (2012b) encoded the contextual information by exploiting the feature descriptors extracted from a focal person and his/her surrounding area. These descriptors are more effective than feature descriptors without contexts (e.g., histogram of oriented gradients (HOG) (Dalal and Triggs, 2005)). However, in the models, the activity of each person is classified independently, and therefore, the spatial and temporal consistency in a group is not always ensured.

In order to obtain this consistency, the question "Which people are in the same group?" must be answered, and an activity in each group must be

optimized. To answer the question, Amer and Todorovic (2011) optimized activities around deformable grids, while Lan et al. (2010b), Choi et al. (2011), Choi and Savarese (2012), Khamis et al. (2012a,b) used graph structures that describe the interactions between persons. However, the models used in these studies cannot describe the "multi-scale" interactions in various features, such as position, size, motion and time sequence, although there exist various types, sizes, and shapes of groups, as shown in Fig. 2. The model proposed by Amer and Todorovic (2011) depended on the density and position of the grids, and therefore, it was difficult to exploit long-range relationships. In the model proposed by Lan et al. (2010b), the person-person interactions were latent and learned automatically; however, their model was restricted to modeling contextual information in a single frame, and was not designed such that temporal consistency was ensured. Choi et al. (2011) and Khamis et al. (2012a,b) considered temporal consistency for a person or group; however, in their model, the person-person interactions that were considered were restricted only in consecutive frames to compute reasonably. The results of these models are likely to be affected by temporary misclassification. Choi and Savarese (2012) exploited a hierarchical model to classify collective activities jointly; however, they assumed there exists only one collective activity in a certain time frame. Therefore, the method cannot model multiple collective activities in a scene, such as that shown in Fig. 2(b), where some persons are waiting at a street intersection, while others are crossing. Considering real-world applications, such as surveillance monitoring, this assumption is not natural.

In contrast, our proposed model describes the "multi-scale" interactions

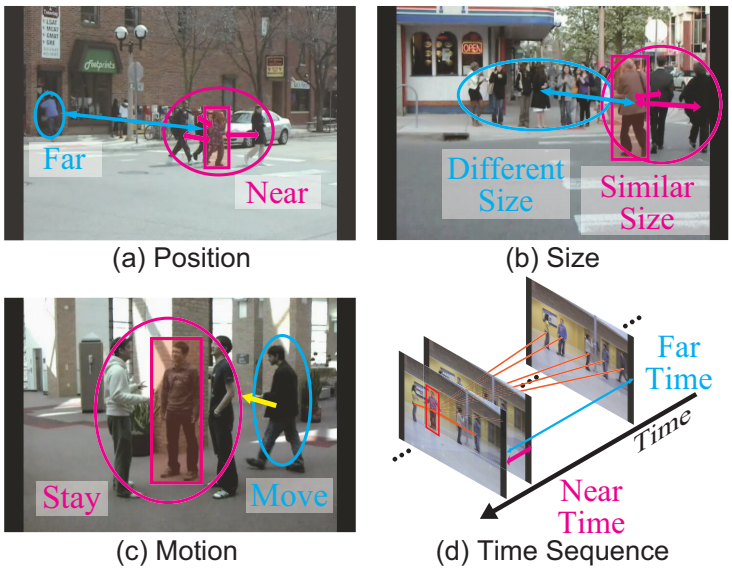

Figure 2: Which people are in the same group? For dividing people into groups, various criteria, such as (a) position, (b) size, (c) motion, and (d) time sequence, can be used.

in various features, i.e., position, size, motion, and time sequence. This means that our model is able not only to describe the long-range relationships among people in both time and space, but also to consider multiple collective activities in a certain time frame. In particular, we use fully connected conditional random fields (CRFs), which consider all the interactions among the people in a video clip, and alter the interaction strength according to the degree of their similarity. This model is able to represent the various features over a "multi-scale" in a single unified model. In general, the calculation cost of a fully connected model is intractable when strict estimation is conducted; however, the cost is reduced to linear in the number of detected persons using a highly efficient approximation method in which the pairwise potentials are modeled using Gaussian kernels (Krähenbühl and Koltun, 2011).

We summarize the main contributions of this paper. (1) We propose a novel method for consistent collective activity recognition in video images using a fully connected model. In the model, we do not restrict the person-person interactions that are considered heuristically, but instead consider all the interactions among the people in a video clip. (2) We describe the person-person interactions over the multi-scale, using various features: position, size, motion, and time sequence. The interaction strength among the people is altered according to the degree of their similarity in the features. (3) We perform the inference with linear complexity in the number of detected persons, using an approximation method in which the pairwise potential are modeled using Gaussian kernels. (4) We evaluate our model on two publicly available datasets. The experimental results show that our fully connected model outperforms other graph structures, such as the unary only model, and the adjacently connected model, as well as state-of-the art models (Choi et al., 2009, 2011; Khamis et al., 2012a,b; Lan et al., 2010a; Kaneko et al., 2012b). Portions of this paper appeared previously in Kaneko et al. (2012a). In this paper, we additionally evaluate our model on the dataset (Choi and Savarese, 2012) and report the results to make our contribution stronger. We also present comparisons with other graph structures, and an additional analysis of the qualitative results, having clarified the characteristic of our model. Moreover, we evaluate a novel combination of our fully connected model and a state-of-the art feature descriptor (Kaneko et al., 2012b), and report the results of a comparison of our proposed model and state-of-the art models.

The rest of the paper is organized as follows. First, in Section 2, we

present our framework of consistent collective activity recognition in video images. Next, in Section 3, the details of learning and inference of the model are given. In Section 4, we report our experimental results quantitatively and qualitatively. Finally, we summarize our paper and present our conclusions in Section 5.

2. Consistent Collective Activity Recognition with Fully Connected CRFs

2.1. Model Overview

The main goal of our study is to ensure the spatial and temporal consistency of the activity in each group in collective activity recognition. For this purpose, our method uses CRFs (Lafferty et al., 2001). CRFs are a probabilistic framework for labeling and segmenting structured data and able to deal with arbitrary dependencies on the observation sequence in a single unified model (He et al., 2004; Shotton et al., 2006). Specifically, in order to handle "multi-scale" interactions, our method uses fully connected CRFs. Instead of specifying the interactions among the people heuristically, our model describes the interactions in position, size, motion, and time sequence as the variable potentials, according to the degree of their similarity, in order to allow various types, sizes, and shapes of groups to be treated.

We now give a brief overview of our model. In the preprocessing step, persons in a video clip are found. Next, features (e.g., HOG (Dalal and Triggs, 2005) and optical flow) are extracted from the detected bounding box. The unary potentials and the pairwise potentials are calculated using the features, and integrated via fully connected CRFs. We present the technical

details of our model in the following sections.

2.2. Model Formulation

Fully Connected CRFs Model. Given a video clip, our method first detects persons using an efficient human detector. In our implementation, we use the approach of Felzenszwalb et al. (2008). The observed data of the detected persons are defined as $\boldsymbol{x} = \{x_1, ..., x_N\}$, where x_i is the observed data of the i-th person and N is the number of detected persons in the video clip. Let the corresponding activity classes be given by $\boldsymbol{y} = \{y_1, ..., y_N\}$. The domain of each variable y_i is a set of activity classes $\mathcal{L} = \{l_1, ..., l_K\}$, where K is the number of activity classes. A conditional random field $(\boldsymbol{x}, \boldsymbol{y})$ is characterized by a Gibbs distribution:

$$P(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{Z(\boldsymbol{x})} \exp(-E(\boldsymbol{y}|\boldsymbol{x})), \tag{1}$$

where $Z(\boldsymbol{x}) = \sum_{\boldsymbol{y}'} \exp(-E(\boldsymbol{y}'|\boldsymbol{x}))$ is the partition function that normalizes the distribution, and $E(\boldsymbol{y}|\boldsymbol{x})$ is the Gibbs energy, which is associated with a configuration $\boldsymbol{y}$ conditioned on $\boldsymbol{x}$. In the fully connected pairwise CRF model, the Gibbs energy is defined as

$$E(\boldsymbol{y}|\boldsymbol{x}) = \sum_i \underbrace{\psi_u(y_i)}_{\text{unary potential}} + \sum_i \sum_{j>i} \underbrace{\psi_p(y_i, y_j)}_{\text{pairwise potential}}, \tag{2}$$

where $\psi_u(y_i)$ is the unary potential and $\psi_p(y_i, y_j)$ is the pairwise potential. For notational convenience, we omit the conditioning in the rest of this paper, and use $\psi_c(\boldsymbol{y}_c)$ to denote $\phi_c(\boldsymbol{y}_c|\boldsymbol{x})$ for each clique c.

Unary Potential. The unary potential $\psi_u(y_i)$ is computed independently for each person by a classifier that produces a distribution over the activity label

y_i given a contextual feature descriptor

$$\psi_u(y_i) = -\log(P(y_i)), \tag{3}$$

where $P(y_i)$ represents the probability that the activity of i-th person is y_i. $P(y_i)$ is calculated by normalizing the classifier scores on the descriptor using a softmax function. In our implementation, we use the descriptors that encode information about not only the action of an individual person, but also the behavior of other people nearby (Choi et al., 2009; Kaneko et al., 2012b). The details are described in Section 2.3.

Pairwise Potential. Since the output of the unary classifier for each person is produced independently of the outputs of the classifiers for other people, the recognition result achieved by the unary classifiers alone is generally noisy and inconsistent. To obtain consistency, we exploit the pairwise potential. The pairwise potential $\psi_p(y_i, y_j)$ represents the interactions between persons. In our fully connected model, the graph structure is the complete graph on $\boldsymbol{y}$, and the pairwise potential is computed for all the sets of persons in a video clip. The detailed explanation of our graph structure is described in Fig 3. In our model, the pairwise potential is defined as

$$\psi_p(y_i, y_j) = \mu(y_i, y_j)k(\boldsymbol{f}_i, \boldsymbol{f}_j), \tag{4}$$

where $\mu(y_i, y_j)$ is the label compatibility function given by the Potts model (Boykov and Jolly, 2001): $\mu(y_i, y_j) = [y_i \neq y_j]$. It introduces a penalty for similar persons that are assigned different labels. The vectors $\boldsymbol{f}_i$ and $\boldsymbol{f}_j$ are feature vectors for the i-th and j-th persons, and $k(\boldsymbol{f}_i, \boldsymbol{f}_j)$ is the Gaussian kernel defined by the positions p_i and p_j, sizes s_i and s_j, motions m_i and m_j,

times t_i and t_j, and weight w:

$$k(\boldsymbol{f}_i, \boldsymbol{f}_j) = w \exp\left(-\frac{|p_i - p_j|^2}{2\theta_1^2} - \frac{|s_i - s_j|^2}{2\theta_2^2} - \frac{|m_i - m_j|^2}{2\theta_3^2} - \frac{|t_i - t_j|^2}{2\theta_4^2}\right). \quad (5)$$

The kernel is inspired by the observation that the persons in the same group have similarities in position, size, motion, and time sequence, as illustrated in Fig 2.

It should be noted that we normalize positions and sizes according to the median size of all the persons in the video clip, in order to describe the interaction strength as a relative rather than an absolute quantity. The motion is computed using different methods according to whether a video clip is captured using a moving or a fixed camera. When using a moving camera, the motion is calculated by subtracting the median optical flow without the bounding boxes from the mean optical flow within the bounding box. The former optical flow represents the camera motion, while the latter optical flow represents the person motion in the image. When using a fixed camera, the motion is defined as the mean optical flow within the bounding box. The optical flow is computed using the approach of Sun et al. (2010).

2.3. Contextual Feature Descriptors

In this section, we describe our method for encoding contextual information into feature descriptors, and calculate the probability in equation (3). We use two descriptors: The *action context (AC) descriptor* proposed by Lan et al. (2010a); and the *combination of the action context and relative action context descriptors (AC-RAC)* that we previously proposed (Kaneko et al.,

2012b). In the experiments, we use the AC descriptor as the baseline to evaluate our model in comparison with other graph topologies. In Kaneko et al. (2012b), it was shown that the AC-RAC descriptor outperforms other previous descriptors (Choi et al., 2009, 2011; Lan et al., 2010a). We integrate our model with this efficacious descriptor and compare it with state-of-the-art methods.

The Action Context Descriptor. The AC descriptor (Lan et al., 2010a) is a per-person descriptor; each descriptor is calculated by concatenating the action descriptor, which captures the action of the focal person, and the context descriptor, which captures the behavior of nearby people.

The action descriptor has a bag-of-words style. We employ the person descriptors (e.g., HOG (Dalal and Triggs, 2005)) as the underlying representation. We then train a multiclass SVM classifier associated with action labels. In the experiments, we use a linear SVM implementation of LIBLINEAR (Fan et al., 2008). Using the score returned by the SVM classifier, the i-th person is represented as the K-dimensional vector $F_i = [S_{1i}, S_{2i}, ..., S_{Ki}]$, where K is the number of action classes, and S_{ki} is the score of classifying the i-th person to the k-th action class.

When the action descriptor has been computed for each person, the context descriptor is calculated by integrating the action descriptor of nearby people in the "context region." The context region is further divided into M regions, called "sub-context regions," in space and time, and then the context descriptor is represented as the $M \times K$ dimensional vector:

$$C_i = [D_{1i}, ..., D_{Mi}]$$

$$= \left[\max_{j \in \mathcal{N}_1(i)} S_{1j}, ..., \max_{j \in \mathcal{N}_1(i)} S_{Kj}, ..., \max_{j \in \mathcal{N}_M(i)} S_{1j}, ..., \max_{j \in \mathcal{N}_M(i)} S_{Kj} \right], \quad (6)$$

where D_{mi} is called the "sub-context descriptor" representing the context in the m-th sub-context region of the i-th person, and $\mathcal{N}_m(i)$ denotes the indices of people in the sub-context region.

The AC descriptor of the i-th person, A_i, is computed by concatenating its action descriptor F_i and its context descriptor C_i: $A_i = [F_i, C_i]$. We then run the multiclass SVM classifier on the AC descriptor associated with activity labels. The classifier scores are normalized using a softmax function, and incorporated as the unary potential in equation (3).

The Relative Action Context Descriptor. The relative action context (RAC) descriptor (Kaneko et al., 2012b) is a refinement of the AC descriptor. Unlike the AC descriptor, the RAC encodes the relative relationship (e.g., if the focal person is facing left and another person is facing right, the relative relationship is defined as facing the opposite direction), and therefore, the descriptor is invariant under a change in the viewpoint, and consistent within the same category of collective activity.

Similarly to Lan et al. (2010b), we define actions by concatenating poses and activities (e.g., talking and facing right). This means that the action descriptor and the sub-context descriptor are $K(= U \times V)$ dimensional vectors, where U is the number of activity classes and V is the number of pose classes. Using U and V, we redefine the action descriptor F_i, and the sub-context descriptor D_{mi} in the AC descriptor:

$$\begin{aligned} F_i &= [S_{1i}, S_{2i}, ..., S_{Ki}] \\ &= [S_{11i}, S_{12i}, ..., S_{uvi}, ..., S_{UVi}], \end{aligned} \quad (7)$$

$$
\begin{aligned}
D_{mi} &= \left[\max_{j\in\mathcal{N}_m(i)} S_{1j}, ..., \max_{j\in\mathcal{N}_m(i)} S_{Kj}\right] \\
&= \left[\max_{j\in\mathcal{N}_m(i)} S_{11j}, \max_{j\in\mathcal{N}_m(i)} S_{12j}, ..., \max_{j\in\mathcal{N}_m(i)} S_{uvj}, ..., \max_{j\in\mathcal{N}_m(i)} S_{UVj}\right]. \quad (8)
\end{aligned}
$$

The RAC descriptor is calculated by shifting the AC descriptor based on the pose of the focal person. First, the pose of the i-th person, $\hat{v}_i$, is calculated from the person descriptor (e.g., HOG (Dalal and Triggs, 2005)) using a multiclass SVM classifier. In terms of pose $\hat{v}_i$, the i-th person's relative action descriptor $\hat{F}_i$ and its relative sub-context descriptor $\hat{D}_{mi}$ are defined as

$$
\begin{aligned}
\hat{F}_i &= \big[S_{1\hat{v}_i i}, ..., S_{1Vi}, S_{11i}, ..., S_{1(\hat{v}_i-1)i}, ..., \\
&\quad S_{U\hat{v}_i i}, ..., S_{UVi}, S_{U1i}, ..., S_{U(\hat{v}_i-1)i}\big], \quad (9) \\
\hat{D}_{mi} &= \left[\max_{j\in\mathcal{N}_m(i)} S_{1\hat{v}_i j}, ..., \max_{j\in\mathcal{N}_m(i)} S_{1Vj}, \max_{j\in\mathcal{N}_m(i)} S_{11j}, \max_{j\in\mathcal{N}_m(i)} S_{1(\hat{v}_i-1)j}, ..., \right. \\
&\quad \left. \max_{j\in\mathcal{N}_m(i)} S_{U\hat{v}_i j}, ..., \max_{j\in\mathcal{N}_m(i)} S_{UVj}, \max_{j\in\mathcal{N}_m(i)} S_{U1j}, \max_{j\in\mathcal{N}_m(i)} S_{U(\hat{v}_i-1)j}\right]. \quad (10)
\end{aligned}
$$

The relative context descriptor of the i-th person, $\hat{C}_i$, is computed by concatenating its relative sub-context descriptor: $\hat{C}_i = [\hat{D}_{1i}, ..., \hat{D}_{Mi}]$. Finally, the RAC descriptor of the i-th person, R_i, is computed by concatenating its relative action descriptor $\hat{F}_i$ and its relative context descriptor $\hat{C}_i$: $R_i = [\hat{F}_i, \hat{C}_i]$.

The Combination of the AC and RAC Descriptors. We now describe the method for combining the AC and RAC descriptors (AC-RAC) (Kaneko et al., 2012b). After extracting the AC and RAC descriptors, we run the multiclass SVM classifier on each of the descriptors associated with activity labels, and transform classifier scores into probabilities via softmax transfor-

mation. We then combine them via the MAX rule (Hatef et al., 1998):

$$\hat{y}_i = \arg\max_{y_i} P_i(y_i) \quad \text{s.t.} \quad P_i(y_i) = \max_k P_i(y_i|d_k), \tag{11}$$

where $P_i(y_i)$ is the probability that the activity of the i-th person is y_i, and $P_i(y_i|d_1)$ and $P_i(y_i|d_2)$ are the probability calculated from the AC and RAC descriptors, respectively. The probability $P_i(y_i)$ is incorporated as the unary potential in equation (3).

It should be noted that, when we use the AC-RAC descriptor in our implementation, we employ two post processes (threshold processing and Gaussian filtering) on the AC and RAC descriptors to accelerate the performance (Kaneko et al., 2012b).

3. Inference and Learning

3.1. Inference

In inference, the maximum a posteriori (MAP) labeling of the random field is estimated:

$$\hat{\boldsymbol{y}} = \arg\max_{\boldsymbol{y}\in\mathcal{L}^N} P(\boldsymbol{y}|\boldsymbol{x}). \tag{12}$$

The exact distribution $P(\boldsymbol{y}|\boldsymbol{x})$ for all the sets of labels $\mathcal{L}^N$ is computationally intractable; however, the calculation cost is reduced to linear in the number of detected persons via the highly efficient approximation method described in Krähenbühl and Koltun (2011), because we define the pairwise potentials in our model as Gaussian kernels. The approximation method uses cross bilateral filtering techniques within a mean field approximation framework. The mean field approximation finds a product of independent

marginals $Q(\boldsymbol{y}) = \prod_i Q_i(y_i)$ close to $P(\boldsymbol{y})$ in terms of minimizing the KL-divergence $\mathbf{D}(Q||P)$ (Koller and Friedman, 2009). By considering the fixed-point equations that hold at the stationary points of KL-divergence, the following iterative update equation is derived for $Q_i(y_i = l)$ given $Q_j(y_j)$ for all $j \neq i$:

$$\begin{aligned} Q_i(y_i = l) &= \frac{1}{Z_i}\Bigg\{-\psi_u(y_i) \\ &\quad - \sum_{l' \in \mathcal{L}} \mu(l, l') \sum_{j \neq i} k(\boldsymbol{f}_i, \boldsymbol{f}_j) Q_j(l')\Bigg\}, \end{aligned} \tag{13}$$

where Z_i normalizes the marginal at node i. A naive implementation of this approximation has quadratic complexity in the number of variables N, when calculating a message passing step: $\tilde{Q}_i(l) = \sum_{j \neq i} k(\boldsymbol{f}_i, \boldsymbol{f}_j) Q_j(l)$. However, in Krähenbühl and Koltun (2011), it is shown that the message passing step can be expressed as a convolution with a Gaussian kernel $\tilde{Q}_i(l) = [G \otimes Q(l)](\boldsymbol{f}_i) - Q_i(l)$, where G is a Gaussian kernel and $\otimes$ is the convolution operator, and it is possible to leverage high-dimensional filtering, such as the permutohedral lattice method (Adams et al., 2010). This reduces the calculation complexity from quadratic to linear in the number of variables N. It should be noted that Krähenbühl and Koltun (2011) report that the inference time is less than one sec for several tens of thousands of variables.

3.2. Learning

In learning, we use piecewise training (Sutton and McCallum, 2005). In piecewise training, the model is divided into several components, each of which is trained independently. Theoretically speaking, piecewise training minimizes an upper bound on the log partition function of the model. The

experimental results presented in Shotton et al. (2006) and Sutton and McCallum (2005) show that piecewise training often performs comparably to global training, even when joint full inference is used. In our model, the unary potentials are trained first, using the contextual feature descriptors described in Section 2.3. Next, we learn the kernel parameters w, θ_1, θ_2, θ_3, and θ_4 in the pairwise potentials. w can be found efficiently by exploiting expectation maximization and high-dimensional filtering. However, it is not easy to optimize the kernel widths θ_1, θ_2, θ_3, and θ_4 globally, due to their non convexity on log-loss criteria. Therefore, we use a grid search on the training set with cross-validation for all the kernel parameters.

4. Experiments

4.1. Datasets and Experimental Setup

We evaluated our model on the Collective Activity Dataset introduced in Choi et al. (2009) and the dataset introduced in Choi and Savarese (2012). Hereafter, we call the former the "Dataset I," and the latter the "Dataset II." These datasets were considered appropriate for our evaluation, since they include activities performed by multiple persons in a natural setting. In most previous studies in the human action recognition field, the proposed algorithms were evaluated on standard benchmark datasets such as the KTH (Blank et al., 2005) and Weizmann (Schuldt et al., 2004) datasets. However, these datasets include a single person performing a specific action, and the video clips in the datasets were recorded in a controlled setting with a small amount of camera motion and a clean background. The Hollywood human action dataset (Laptev et al., 2008) and UT-Interaction dataset (Ryoo and

Aggarwal, 2010) are more challenging and contain actions performed by more than one actor; however, the actions (e.g., hand shaking, hugging) are not collective, but rather two persons perform one action together.

The Dataset I is composed of 44 short video clips, including five activity classes: *crossing*, *waiting*, *queuing*, *walking*, and *talking*. The video clips in the dataset were recorded using low resolution hand-held cameras under realistic conditions, including camera shaking, background clutter, and transient mutual occlusions of persons. Some video clips include multiple collective activities in a scene or activity transition. All the persons in every tenth frame of the videos are labeled with the ground truth: pose, activity, and bounding box information.

The Dataset II is composed of 32 short video clips, including six activity classes: *gathering*, *talking*, *dismissal*, *walking together*, *chasing*, and *queuing*. The video clips in the dataset were recorded using a fixed camera in the outdoors. We omitted the activity *queuing* and used the remaining 30 short video clips for evaluation, because the number of video clips containing *queuing* is only two, and therefore, too small to evaluate using a leave-one-video-out cross-validation scheme. It should be noted that each of the other activities is contained in more than five video clips, respectively. While the pose label and bounding box information are annotated per person, the activity label is annotated per frame, since Choi and Savarese (2012) assumed there exists only one collective activity in a certain time frame, and their goal was to classify the activity per frame. However, this assumption is not always natural, because some video clips contain multiple activities in a single frame, e.g., one person is passing by, while the others are chasing. We therefore added

the *walking alone* activity class and re-annotated each person with six activities: *gathering*, *talking*, *dismissal*, *walking together*, *chasing*, and *walking alone*.

We evaluated our model in a way similar to that of Choi et al. (2009, 2011), Khamis et al. (2012a,b), Lan et al. (2010a), and Kaneko et al. (2012b). For each dataset, we used the leave-one-video-out cross-validation scheme. This means that when we classified activities in one video, all the other videos in the dataset were used for training and validation. We report activity classification results on a per-person basis. It should be noted that, in some previous studies (Lan et al., 2010b; Choi and Savarese, 2012), it was assumed that there exists only one collective activity in a certain time frame, and activity classification results were reported on a per-frame basis. This experimental protocol is inadequate to evaluate our model, because our goal was to segment collective activities in a scene where multiple groups exist, as shown in Fig. 2.

The experiments were conducted on an Intel Core i7 processor clocked at 2.2 GHz. The calculation complexity of inference using our fully connected model was $O(N)$, where N (the number of detected persons) was typically in the order of tens, hundreds, or thousands, and inference time was less than 0.1 sec per video sequence for both the datasets. It should be noted that, in Choi and Savarese (2012), the classification and target association take about 1 min per video sequence, given tracklets and observations.

4.2. Evaluation of Graph Structures

In order to evaluate the performance of the proposed model comprehensively, we compared it with several baseline models. The first baselines use

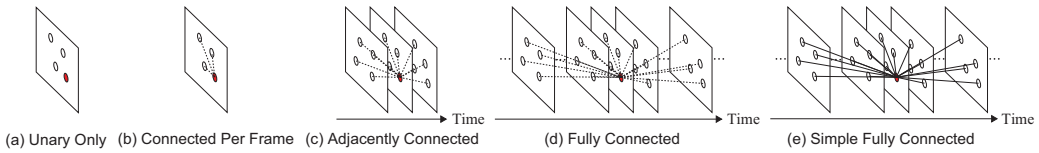

Figure 3: A structure of person-person interactions in each graph model. Each node represents a person in a video. Dashed lines represent that the interaction strength between persons is altered according to the degree of their similarity, while solid lines represent that the interaction strength between persons is constant, regardless of the degree of their similarity. (a) The unary only model (no connection between any pair of nodes); (b) nodes are connected in a frame; (c) nodes are connected in adjacent frames; (d) and (e) all the nodes are connected in a video.

various ways of setting the structures of the person-person interactions. The structures that we considered are shown in Fig. 3(a)-(c), including (a) the *Unary Only* model (no pairwise connection); (b) the graph obtained by connecting persons in a frame (the *Connected Per Frame* model); (c) the graph obtained by connecting persons in adjacent frames (the *Adjacently Connected* model). It should be noted that in our proposed *Fully Connected* model, all the persons in a video are connected, as shown in Fig. 3(d). In the three models shown in Fig. 3(b)-(d), pairwise potentials are defined as Gaussian edge potentials, and therefore, the interaction strength between persons is altered according to the degree of their similarity.

In order to evaluate the Gaussian edge potentials, we also compared the proposed model with the other baseline (which we call *Simple Fully Connected*), fully connected CRFs with constant edge potentials, as shown in Fig. 3(e). In the model, the pairwise potential in equation (2) is defined as $\phi_p(y_i, y_j) = w\mu(y_i, y_j)$. Unlike in the proposed model, the interaction

strength between persons in this model is constant over a video. In order to analyze the efficacy of the graph structure models, we used the AC descriptor (Lan et al., 2010a) as the baseline, and calculated the unary potential in equation (3) based on this descriptor.

Table 1: Comparison of activity classification accuracy levels for different graph structures. Each graph structure has a different structure of person-person interactions or different interaction strength between persons. The details are illustrated in Fig. 3.

Method/Dataset	Dataset I (Choi et al., 2009)	Dataset II (Choi and Savarese, 2012)
Unary Only	67.4%	56.2%
Connected Per Frame	68.6%	56.7%
Adjacency Connected	69.6%	57.6%
Fully Connected	**72.2%**	**59.3%**
Simple Fully Connected	68.7%	55.5%

Quantitative Results. The comparison results of our approach and the baselines are summarized in Table 1. The results shows that our model (*Fully Connected*) outperforms the baselines on the two datasets. The *Simple Fully Connected* model has a fully connected graph structure, as does the *Fully Connected* model; however, the former is significantly outperformed by the latter, since the former exploits the long-range interaction but cannot deal with multiple collective activities and activity transition. The confusion matrices of our model and the baseline (the *Unary Only* model) are illustrated in Fig 4. These confusion matrices show that our proposed *Fully Connected* model achieves a significant improvement over the baseline. It should be

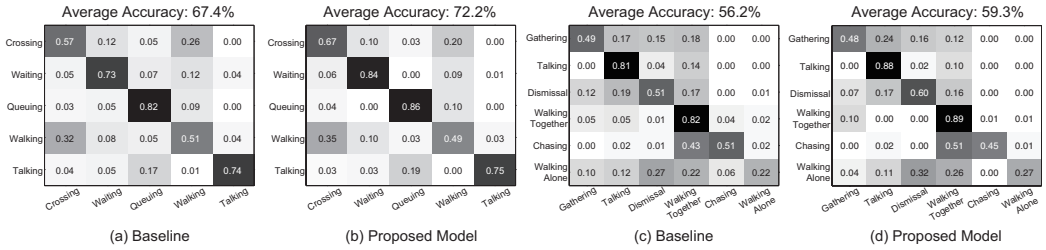

Figure 4: Confusion matrices for activity classification on the two datasets. (a) and (b) show the confusion matrices for collective activity using the baseline (the *Unary Only* model) and proposed method on the Dataset I (Choi et al., 2009), respectively. (c) and (d) compare the two methods on the Dataset II (Choi and Savarese, 2012). In both cases, our fully connected model significantly outperforms the baseline method. In a confusion matrix, rows represent ground truth and columns represent prediction. Each row is normalized to sum to 1.

noted that *walking* vs *crossing* for the Dataset I is still ambiguous in our model, because whether a person is *walking* or *crossing* often depends not on the person-person interactions but on the environmental setting, such as a sidewalk or a pedestrian crossing. In both our model and the baseline, *walking alone* on the Dataset II is confused with *dismissal* and *walking together*, because *walking alone* is a single person activity and the neighbors of the person have a variety of appearances, e.g., one person passing by other people often seems like dismissal. However, the number of confusions is reduced using our model.

Qualitative Results. We also visualize the qualitative results using different structures of person-person interactions in Figs. 5–7. Fig. 5 shows typical success and failure examples, Fig. 6 shows examples of the scenes where multiple groups exist, and Fig. 7 shows examples of the case where the activities

transition from one to another.

Fig. 5 shows typical success and failure examples for the Dataset I. Fig. 5(a)–(b) demonstrates that the classification result achieved by the *Unary Only* model is often noisy and inconsistent, since the activity of each person is classified independently in the model. In the *Connected Per Frame* model and the *Adjacently Connected* model, collective activities are optimized in a frame and consecutive frames, respectively; however, misclassification of some persons in the duration of the video clip often causes misclassification of others. In contrast, in our *Fully Connected* model, the misclassification is fixed, and we obtain the temporal and spatial consistency of the activity in a group by leveraging the cues over the multi-scale. However, the wrong classification in the *Unary Only* model causes any graph structure to classify everyone incorrectly, as shown in Fig. 5(c).

Fig. 6 visualizes examples of the scenes where multiple groups exist in the Dataset I. Fig. 6(a)–(b) shows that the *Fully Connected* model yields misclassification errors, but achieves consistency in each group, while the other models fail to obtain consistency. This is because the *Connected Per Frame* model and the *Adjacently Connected* model are likely to be influenced by the misclassification of some persons in the same frame and consecutive frames, respectively. However, the high level of confidence in the wrong label in the *Unary Only* model causes any structure to fail to obtain consistency, as shown in Fig. 6(c).

Fig. 7 visualizes examples of the case where the activities transition from one to another in the Dataset II. In particular, the activity transitions from *gathering* to *talking*. Since the point at which *gathering* changes to *talking* is

ambiguous, that of the *Fully Connected* model is not completely consistent with that of the ground truth; however, only our model succeeds in recognizing the transition between the two activities, in contrast to any other model.

Figure 5: (Best viewed in color) Visualization of typical success and failure examples for the Dataset I (Choi et al., 2009) using different structures of person-person interactions. The labels C (magenta), S (blue), Q (cyan), W (red), T (green), and NA (white) indicate *crossing*, *waiting*, *queuing*, *walking*, *talking*, and *not assigned*, respectively. The first row shows the ground truth; the second row shows results using the *Unary Only* model; the third row shows results using the *Connected Per Frame* model; the fourth row shows results using the *Adjacently Connected* model; the fifth row shows results using our proposed *Fully Connected* model. The first three columns (a) represent results in the consecutive frames in the same video clip, and the last two columns (b)-(c) represent results in other video clips. In the first four columns (a)-(b), temporary misclassification in the *Unary Only* model is fixed in the *Fully Connected* model. In the fifth column (c), any structure fails to classify collective activities, affected by the wide and continuous misclassification in the *Unary Only* model.

Figure 6: (Best viewed in color) Visualization of examples of the scenes where multiple groups exist in the Dataset I (Choi et al., 2009). The labels C (magenta), S (blue), Q (cyan), W (red), T (green), and NA (white) indicate *crossing*, *waiting*, *queuing*, *walking*, *talking*, and *not assigned*, respectively. The first row shows the ground truth; the second row shows results using the *Unary Only* model; the third row shows results using the *Connected Per Frame* model; the fourth row shows results using the *Adjacently Connected* model; the fifth row shows results using our proposed *Fully Connected* model. The first three columns (a) represent results in the consecutive frames in the same video clip, and the last two columns (b)-(c) represent results in other video clips. In the first four columns (a)-(b), the *Fully Connected* model achieves consistency in a group, although the other models fail to obtain consistency in some scenes. In the fifth column (c), the high level of confidence in the wrong label in the Unary Only model causes any structure to fail to obtain consistency.

Figure 7: (Best viewed in color) Visualization of examples of the case where the activities transition from one to another in the Dataset II (Choi and Savarese, 2012). In particular, the activity transitions from *gathering* to *talking*. The labels G (magenta), T (green), D (cyan), and W (red) indicate *gathering*, *talking*, *dismissal*, and *walking together*, respectively. The first row shows the ground truth; the second row shows results using the *Unary Only* model; the third row shows results using the *Connected Per Frame* model; the fourth row shows results using the *Adjacently Connected* model; the fifth row shows results using our proposed *Fully Connected* model. Since the point at which *gathering* changes to *talking* is ambiguous, that of the *Fully Connected* model is not completely consistent with that of the ground truth; however, only our model succeeds in recognizing the transition between the two activities, in contrast to any other model.

4.3. Comparison with State-of-the-Art Methods

We also compared our method with recent methods (Choi et al., 2009, 2011; Khamis et al., 2012a,b; Lan et al., 2010a; Kaneko et al., 2012b). So that the comparison would be fair, we used the same leave-one-video-out scheme described in the studies; we report activity classification results on a per-person basis. In order to evaluate our fully connected model in comparison with state-of-the art models, we calculated the unary potentials in equation (3) using not only the AC descriptor (Lan et al., 2010a) but also the combination of the AC and RAC descriptors (AC-RAC) (Kaneko et al., 2012b).

Results. The comparison results of the activity classification accuracy levels for different methods for the Dataset I and Dataset II are summarized in Table 2. We analyzed our results mainly on the Dataset I, because in no previous study a model was evaluated using the same protocol as ours in the Dataset II.

In the models using the AC descriptor as the baseline, our model (*AC + FC-CRF*) outperforms the other models: *AC + Frame Cues* (Khamis et al., 2012a), *AC + Track Cues* (Khamis et al., 2012b) and *AC + Frame/Track Cues* (Khamis et al., 2012a), although our baseline (*AC*) is inferior to those of (Khamis et al., 2012a,b). It should be noted that *AC + Frame/Track Cues* (Khamis et al., 2012a) connects only people with overlapping bounding boxes in consecutive frames to improve the inference speed, while our model can solve the fully connected model in less than 0.1 sec. *AC + FC-CRF* also outperforms *STV + MC* (Choi et al., 2009) and *RSTV + MRF* (Choi et al., 2011). It should be noted that these models (Choi et al., 2009, 2011) em-

ploy additional trajectory information of each person to obtain consistency. However, it is not easy to obtain the correct trajectory in a scene where transient mutual occlusions of persons exist. Mistakes produced during the tracking step can influence the performance of the model during recognition. Khamis et al. (2012a,b) also did not use the trajectory, and instead solved the identity maintenance problem and action recognition simultaneously.

We also show the results obtained using our previously proposed descriptor (*AC-RAC*) (Kaneko et al., 2012b) and integration of our model and the descriptor (*AC-RAC + FC-CRF*). In the two datasets, *AC-RAC* outperforms the other feature descriptors, and the integration of our model and the descriptor (*AC-RAC + FC-CRF*) achieves the best performance. Choi and Savarese (2012) reported an accuracy (74.4%) that is competitive with that achieved by our method (74.7%) on the Dataset I. However, since their accuracy level was achieved using a different experimental protocol, i.e., they assigned the per-scene collective activity labels that they obtained with four-fold experiments to each individual, it is not directly comparable to the accuracy levels listed in Table 2.

In the Dataset II, no previous method was directly comparable to our method, because, in no previous study, a model was evaluated on a *per-person* basis like ours, while Choi and Savarese (2012) evaluated their model on a *per-frame* basis, and reported an accuracy (74.3% mean-per-class using baseline method and 79.2% mean-per-class using their proposed method). The direct comparison between our method and their method using the same experimental protocol would be an interesting topic, however, it would be hard to perform the direct comparison, because the assumptions of our and their

models are basically different and it is required to reformulate the problems significantly. Our model aims to recognize an activity in each group even in the scene where multiple activity groups exist, while their model aims to recognize one main collective activity per frame under the assumption that there is only one collective activity per scene per time stamp. It should be noted that our and their model require the different annotated information in training process. Our model requires the activity label annotated *per person*, while their model requires the activity label annotated *per frame*. In the scene where some activity groups exist, they assigned each frame into one activity label by taking the majority of activities of persons in that frame.

Table 2: Comparison of activity classification accuracy levels for different methods on the Dataset I (Choi et al., 2009) and Dataset II (Choi and Savarese, 2012). The methods without a plus sign (+) do not use graph structure, s.t., unary only models, such as shown in Fig. 3(a), while the methods with a plus sign (+) use a graph structure, such as shown in Fig. 3(b)–(e). The last four rows show the results obtained using our implementation. In the two datasets, our previously proposed descriptor (*AC-RAC*) (Kaneko et al., 2012b) outperforms the other feature descriptors, and the integration of our model and the descriptor (*AC-RAC + FC-CRF*) achieves the best performance.

Method/Dataset	Dataset I (Choi et al., 2009)	Dataset II (Choi and Savarese, 2012)
AC (Lan et al., 2010a)	68.2%	-
STV (Choi et al., 2009)	64.3%	-
STV + MC (Choi et al., 2009)	65.9%	-
RSTV (Choi et al., 2011)	67.2%	-
RSTV + MRF (Choi et al., 2011)	70.9%	-
AC (Khamis et al., 2012b)	68.8%	-
AC + Track Cues (Khamis et al., 2012b)	70.9%	-
AC + Frame Cues (Khamis et al., 2012a)	70.7%	-
AC + Frame/Track Cues (Khamis et al., 2012a)	72.0%	-
AC	67.4%	56.2%
AC + FC-CRF	**72.2%**	**59.3%**
AC-RAC	**71.9%**	**66.4%**
AC-RAC + FC-CRF	**74.7%**	**70.7%**

5. Conclusion and Discussion

In this paper, we described a method for consistent collective activity recognition that uses fully connected CRFs, which consider all the interactions among the people in a video clip and alter the interaction strength according to the degree of their similarity. Our model leverages various features, such as position, size, motion, and time sequence, over a "multi-scale"

in a single unified model, in order to allow various types, sizes, and shapes of groups to be treated. The intractability of the fully connected model is overcome by describing the pairwise potentials as Gaussian kernels and using a highly efficient approximation method. Our experimental results showed that our model is robust against temporary misclassification and able to deal with multiple activities in a scene and activity transition. Quantitative evaluation on the two publicly available datasets showed that our fully connected model outperforms state-of-the art models, as well as other graph structures, such as the unary only model and the adjacently connected model.

Future directions include investigation of other useful contexts for collective activity recognition, such as environmental settings; extension to online processing for online applications; and research on efficient learning techniques as a substitute for grid search to optimize the kernel parameters efficiently.

References

Adams, A., Baek, J., Davis, M. A., 2010. Fast high-dimensional filtering using the permutohedral lattice. In: Computer Graphics Forum.

Amer, M. R., Todorovic, S., 2011. A chains model for localizing participants of group activities in videos. In: International Conference on Computer Vision (ICCV).

Blank, M., Gorelick, L., Shechtman, E., Irani, M., Basri, R., 2005. Actions as space-time shapes. In: International Conference on Computer Vision (ICCV).

Boykov, Y. Y., Jolly, M.-P., 2001. Interactive graph cuts for optimal boundary and region segmentation of objects in n-d images. In: International Conference on Computer Vision (ICCV).

Choi, W., Savarese, S., 2012. A unified framework for multi-target tracking and collective activity recognition. In: European Conference on Computer Vision (ECCV).

Choi, W., Shahid, K., Savarese, S., 2009. What are they doing?: Collective activity classification using spatio-temporal relationship among people. In: International Workshop on Visual Surveillance (VS).

Choi, W., Shahid, K., Savarese, S., 2011. Learning context for collective activity recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Dalal, N., Triggs, B., 2005. Histograms of oriented gradients for human detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Fan, R. E., Chang, K. W., Hsieh, C. J., Wang, X. R., Lin, C. J., 2008. LIBLINEAR: A library for large linear classification. Journal of Machine Learning Research (JMLR) 9, 1871–1874.

Felzenszwalb, P., McAllester, D., Ramanan, D., 2008. A discriminatively trained, multiscale, deformable part model. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Hatef, M., Duin, R. P., Matas, J., 1998. On combining classifiers. IEEE

Transaction on Pattern Recognition and Machine Intelligence (PAMI) 20, 226–239.

He, X., Zemel, R. S., Carreira-Perpiñán, M. A., 2004. Multiscale conditional random fields for image labeling. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Kaneko, T., Shimosaka, M., Odashima, S., Fukui, R., Sato, T., 2012a. Consistent collective activity recognition with fully connected CRFs. In: International Conference on Pattern Recognition (ICPR).

Kaneko, T., Shimosaka, M., Odashima, S., Fukui, R., Sato, T., 2012b. Viewpoint invariant collective activity recognition with relative action context. In: Workshop on Action Recogniton and Pose Estimation in Still Images (APSI).

Khamis, S., Morariu, V. I., Davis, L. S., 2012a. Combining per-frame and per-track cues for multi-person action recognition. In: European Conference on Computer Vision (ECCV).

Khamis, S., Morariu, V. I., Davis, L. S., 2012b. A flow model for joint action recognition and identity maintenance. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Koller, D., Friedman, N., 2009. Probabilistic Graphical Models: Principles and Techniques. MIT Press.

Krähenbühl, P., Koltun, V., 2011. Efficient inference in fully connected CRFs with Gaussian edge potentials. In: Advances in Neural Information Processing Systems (NIPS).

555 Lafferty, J., McCallum, A., Pereira, F., 2001. Conditional random fields:
556 Probabilistic models for segmenting and labeling sequence data. In: Inter-
557 national Conference on Machine Learning (ICML).

558 Lan, T., Wang, Y., Mori, G., 2010a. Retrieving actions in group contexts.
559 In: International Workshop on Sign Gesture Activity (SGA).

560 Lan, T., Wang, Y., Yang, W., Mori, G., 2010b. Beyond actions: Discrim-
561 inative models for contextual group activities. In: Advances in Neural
562 Information Processing Systems (NIPS).

563 Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B., 2008. Learning realistic
564 human actions from movies. In: IEEE Conference on Computer Vision and
565 Pattern Recognition (CVPR).

566 Niebles, J. C., Wang, H., Fei-Fei, L., 2006. Unsupervised learning of human
567 action categories using spatial-temporal words. In: British Machine Vision
568 Conference (BMVC).

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

572 Schuldt, C., Laptev, I., Caputo, B., 2004. Recognizing human actions: A
573 local SVM approach. In: International Conference on Pattern Recognition
574 (ICPR).

575 Shotton, J., Winn, J., Rother, C., Criminisi, A., 2006. Textonboost: Joint
576 appearance, shape and context modeling for multi-class object recognition
577 and segmentation. In: European Conference on Computer Vision (ECCV).

Sun, D., Roth, S., Black, M. J., 2010. Secrets of optical flow estimation and their principles. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Sutton, C., McCallum, A., 2005. Piecewise training for undirected models. In: Conference on Uncertainty in Artificial Intelligence (UAI).