STRUCTURED LEARNING FOR CROWD MOTION SEGMENTATION

Habib Ullah and Nicola Conci

DISI - University of Trento - Italy

ABSTRACT

In this paper we present a novel method for motion segmentation in crowded scenes, based on statistical modeling for structured prediction using a Conditional Random Field (CRF). As opposed to other conditional Markov models, CRF overcomes the label bias problem, making it suitable for crowd motion analysis. In our method, a grid of particles is initialized on the scene, and advected using optical flow. The particles are exploited to extract motion patterns, used as input priors for CRF training. Furthermore, we exploit min cut/max flow algorithm to remove the residual noise and highlight the main directions of crowd motion. The experimental evaluation is conducted on a set of benchmark video sequences, commonly used for crowd motion analysis, and the obtained results are compared against other state of the art techniques.

Index Terms— Optical flow, conditional random fields

1. INTRODUCTION

The automatic analysis of activities occurring in crowded scenes is escalating more attention from the research community due to the request for autonomous analysis tools able to overcome the limitations of traditional tracking algorithms in presence of severe occlusions. In this context, object tracking is likely to fail, since tracking of each single subject becomes impracticable [1] [2] [3]. Therefore understanding crowd motion is an open problem by virtue of the potential complexities of tracking individuals and also the vast diversity of the scene under investigation when scene context and crowd dynamics may change considerably over time. It is therefore more appropriate to consider the crowd as a single entity, discarding the information about individuals.

The literature in the area of crowd motion analysis is quite rich, and an overview about earlier algorithms in the area and related issues is presented by Zhan et al. [4]. More recently, Solmaz et al. [5] proposed a method for identifying five crowd behaviors using time integration of the dynamical system defined by the optical flow. Krausz and Bauchhage [6] detected dangerous crowded situation, through the optical flow information. Butenuth et al. [7] proposed an interdisciplinary framework for the analysis of the crowd, which integrates benefits from other techniques. Raghavendra et al. [8] employed Particle Swarm Optimization for global anomaly detection in crowded scenes. The approach can be targeted at real-time applications. In [9], abnormal events are detected in terms of escape panics. For this purpose, the social force model (SFM) [10] is exploited. After the superposition of a fixed grid of particles on each frame, the SFM is used to estimate the interaction forces for describing the crowd behavior. Wu et al. [11] use chaotic invariants of Lagrangian Particle Trajectories to model abnormal patterns in crowded scenes. Ali and Shah [12] exploited the Finite Time Lyapunov Exponent [13] to extract boundaries between different flow regions in the crowd. However, when the optical flow is not precise enough, the boundaries are discontinuous, resulting in oversegmentation in low-density crowd scenes. A more recent related work [14] proposed streaklines for crowd motion segmentation. The streaklines are modeled using a linear dynamical model, but are incapable to encapsulate the crowd dynamics, thus failing to group pixels with common motion patterns. In addition, streaklines cannot capture temporal changes, exhibiting choppy motion segmentation in high density crowd scenes. Another related work [15] uses $alpha$-expansion moves based on graph cuts [16], commonly used to compute a local minimum, for crowd segmentation. There is a great diversity with approaches to crowd segmentation. In general, it is quite difficult to compare different solutions, also due to the lack of ground truth to compare against. Furthermore, different representations of crowd motion and crowd density are combined with different models for crowd segmentation, which are typically tailored to the type of video analyzed, or to a specific scene domain.

In this work we propose to address the problems mentioned above, by training a conditional random field (CRF) to segment the motion flow. We first position a grid of particles over the frame tracked using the Lucas-Kanade optical flow. By tracking the particles, we extract motion patterns, which are used as a-priori information for CRF training. Training is performed by means of the gradient ascent algorithm, so as to maximize the conditional likelihood. Furthermore, the parameters after training are used for CRF to segment the crowd flow in terms of motion directions. In fact, compared to other approaches, such as Hidden Markov Model (HMM), CRF is able to model dense and correlated flow features of crowd since it models the conditional probability allowing relaxation of the strong independence assumptions made by the HMM.
2. CROWD MOTION SEGMENTATION

The method we propose is staged in three main blocks namely: particle advection, CRF inferencing, and refinement of the motion map using graph cut. During the first stage, a grid of particles is disposed on the image plane. Each particle represents a block of pixels of predefined size. Motion patterns, defined in terms of orientation features, are extracted by tracking the particles using the pyramidal Lucas-Kanade optical flow [17]. During this first stage, the orientation features act as a sequential data for inferencing the CRF, resulting into a flow map. The orientations features with the corresponding label sequence are used to learn the CRF parameters during the training stage, and the crowd motion directions are inferred on the test samples. In order to provide a more consistent representation of the crowd motion in the second stage, graph cut [16] is used to filter out the residual noise.

2.1. Inferencing

After positioning the grid of particles over the frame, and tracking it by the Lucas-Kanade optical flow, the orientation features of each particle in term of angle of motion are extracted at regular intervals of $K$ frames. The collected orientation features are stored to construct a feature vector for each particle. The objective of this processing stage is to isolate and filter out the orientation features that would be possible if considered singularly, but that do not contribute to the identification of the crowd motion direction.

A Conditional Random Field (CRF) is a discriminative model used for labeling sequential data. It specifies the probability of a particular label sequence, given the observation sequence. Specifically, $x$ is our input sequence, consisting in $N$ observations collected within the $K$ frames window (i.e. $x = x_1, x_2, \ldots, x_N$), containing the orientation features. Given the observation sequence, the CRF thus signals the most probable label in terms of direction, inferring the output label $y$ ($y = y_1, y_2, \ldots, y_M$) of the respective crowd motion direction, and quantized in $M$ possible values.

$$p(y/x; w) = \frac{\exp \sum_j w_j F_j(x, y)}{Z(x, w)}$$

In Eq. (1), $F_j(x, y)$ is a feature function, which consists of the paired mapping $F_j : X \times Y \to \mathbb{R}$. Each feature function renders the score for any output label $y$ in terms of its relevance to the input observation vector $x$. The flow of inference process is shown in Fig. 1 where $N$ represents the total number of particles tracked. The denominator in Eq. (1) is a partition function $Z(x, w)$, which ranges over all the label set $y$.

$$Z(x, w) = \sum_y \exp \left\{ \sum_j w_j F_j(x, y) \right\}$$

Hence, the partition function acts as a normalization factor. Given orientation features $x$, the corresponding label is obtained as:

$$\hat{y} = \text{argmax}_y p(y/x; w) = \text{argmax}_y \sum_j w_j F_j(x, y) \quad (3)$$

For each $j$, we will obtain different $F_j$ functions, according to the parameter $w_j$ and the test observation sequence $x$. Our main contention in obtaining the probability score for each label sequence is that it is easy to reveal the most probable direction for each particle, which can segment the crowd motion as the scene dynamically changes over time.

2.2. Training

The goal of the training phase is to choose the appropriate values for the parameters $w_j$, so as to maximize the conditional probability of the training examples. For this purpose we exploit the stochastic gradient ascent to maximize the conditional log-likelihood (CLL) of the set of training examples:

$$\frac{\partial}{\partial w_j} \log p(y/x; w) = F_j(x, y) - \frac{\partial}{\partial w_j} \log Z(x, w) \quad (4)$$

For each $w_j$, the partial derivative of CLL is evaluated for single training sequences, i.e., one weight for each feature function $F_j$. The partial derivative with respect to $w_j$ corresponds to the $i$-th value of the feature function for its true label $y$, minus the averaged values of the feature function for all possible labels $y$. Therefore, Eq. (4) can be rewritten as:
\[ \frac{\partial}{\partial w_j} \log p(\mathbf{y}/\mathbf{x}; w) = F_j(\mathbf{x}, \mathbf{y}) - \sum_{\mathbf{y}'} p(\mathbf{y}'/\mathbf{x}; w) \left[ F_j(\mathbf{x}, \mathbf{y}') \right] \]

In order to maximize the conditional log-likelihood by stochastic gradient ascent, \( w_j \) is updated according to Eq. (6) where \( \alpha \) is the learning rate.

\[ w_j = w_j + \alpha(F_j(\mathbf{x}, \mathbf{y}) - \sum_{\mathbf{y}'} p(\mathbf{y}'/\mathbf{x}; w) \left[ F_j(\mathbf{x}, \mathbf{y}') \right]) \]

### 2.3. Sanitizing the segmented motion map

Although the output of the CRF inference is in general quite accurate in indicating the motion flow, it still includes a non negligible amount of noise. In order to remove this spurious data and to better highlight the main motion directions of the crowd flow, we used the \( \alpha \)-expansion moves based on graph cuts [16], which produce a solution within a known factor of the global minimum of the energy function. The minimization process takes place according to Eq. (7)

\[ E(L) = \sum_{p \in P} D_p(L_p) + \sum_{(p,q) \in N} V_{p,q}(L_p, L_q) \]

where \( D_p \) is the so-called data cost term, and \( V_{p,q} \) is the smooth cost term. The \( \alpha \)-expansion minimizes the energy function for a set of labels under the class of smoothness term, called metric. Unlike [15], we exploited both the data cost term and the smooth cost terms so that the resulting labeling fit to the data and accomplishes the desired smoothing. In our implementation, the frame acquired in the preceding stage is divided into blocks of fixed size. Each block is labeled in the frame according to the \( M \) possible labels. The data cost gives different weights to each orientation according to the distance. The higher the distance between the orientations, the higher the data cost. For every node of the analyzed graph we extract the orientation information. This orientation is then compared with all labels in the set, searching in a neighborhood defined a priori. Given \( M \) labels we calculate the data cost term according to Eq. (8) as in [15]:

\[ D(\phi_l) = \min \left( \left| \frac{\phi_l}{P} - \frac{\phi}{P} \right|, M - \left| \frac{\phi_l}{P} - \frac{\phi}{P} \right| \right) \]

where \( P \) is the minimum distance between the labels, \( \phi_l \) is the \( l \)-th orientation in the label set, and \( \phi \) is the current orientation of the element under observation: \( \phi \) is compared against all the labels in the set and the one minimizing the energy function in Eq. (7) is chosen. In order to consolidate the output, the smooth cost term is defined as:

\[ S(\phi_l) = \text{abs}(\phi_l - \phi_{l+1}) \]

As shown in Fig. 2, the \( \alpha \)-expansion moves demonstrate a very good capability in suppressing the residual noise left by the preceding processing stages.

### 3. EXPERIMENTS AND RESULTS

In this section we present the results of our approach outlined in Section 2. For testing purpose we have used six publicly available benchmark video sequences [12] [15] [14] to thoroughly evaluate the effectiveness of our proposed approach. In Fig. 3 and Fig. 4, the first rows present the snapshots of the original video sequences, while the second, third, fourth and fifth rows show the results obtained using (i) pure optical flow, (ii) the method in [15], (iii) streaklines approach [14], and (iv) the proposed approach, respectively.

To neglect regions without motion, we discard small magnitude optical flow. For the extraction of the orientation features for each particle, the resolution of the grid is kept half of the resolution of the video frame. For each particle, the orientation features consist of a vector of \( N = 4 \) observations, where each element of the vector corresponds to the orientation information extracted after each \( K = 8 \) frames. The possible output directions are \( M = 8 \), one label every \( 45^\circ \).

When applying the graph cut, each frame processed by the CRF is divided into blocks \( 2 \times 2 \) pixels. Each block is considered as a single element and scanning is carried out from top-left to bottom-right. For each central block, the spatial neighborhood is set to \( 5 \times 5 \) blocks. For the training phase, we used 800 examples extracted from different video sequences. Each training example is extracted randomly, so that the trained model reflects a relevant and accurate representation of the training data.

The results obtained are divided into categories of human motion in Fig. 3 and traffic motion in Fig. 4. We compare the results obtained with pure optical flow as a baseline. We track the particles for 8 consecutive frames by using the
Lucas-Kanade optical flow. Then, the obtained \textit{tracklets} are drawn according to the selected eight possible output directions. It is evident from the segmentation map, that the simple optical flow representation is not powerful enough to segment the crowd motion. Also, when comparing with the method presented in [15], we can notice inconsistencies in the crowd motion in Fig. 3, and this is evident especially in the first and second video sequences in the first row, where the crowd is moving in semi-circle direction, and from top to bottom on the bridge, respectively.

In Fig. 3, streaklines [14] exhibit choppy motion segmentation for the second and third video sequences, whereas our approach is spatially and temporally pronounced and more accurate. Furthermore, the approach in [15] presents quite unreliable results for the first and third video sequences in the first row in Fig. 4. In fact, it is mainly based on spatial correlation, discarding the temporal component, which turns out to be a discriminant factor. Streaklines [14] create stilted time lag and cannot detect local spatial changes, hence leave spatial crevices in flow and rapid transition between frames as it is evident for three video sequences in Fig. 4.

4. CONCLUSION

We present an approach for segmenting motion in crowded scenes using CRF. For this purpose, we extracted the orientation features by exploiting the optical flow evaluated on a set of particles uniformly distributed on the image plane. The orientation features are used as a-priori to train the CRF. To further clean up the segmented motion map, graph cut is used to consolidate the CRF output. The method demonstrated promising performance when compared with other related works.
5. REFERENCES


