MOTION PATTERN ANALYSIS IN CROWDED SCENES BASED ON HYBRID GENERATIVE-DISCRIMINATIVE FEATURE MAPS

Chongjing Wang, Xu Zhao, Zhe Wu and Yuncai Liu
Department of Automation and Key Laboratory of China MOE for System Control and Information Processing
Shanghai Jiao Tong University

ABSTRACT

Crowded scene analysis is becoming increasingly popular in computer vision field. In this paper, we propose a novel approach to analyze motion patterns by clustering the hybrid generative-discriminative feature maps using unsupervised hierarchical clustering algorithm. The hybrid generative-discriminative feature maps are derived by posterior divergence based on the tracklets which are captured by tracking dense points with three effective rules. The feature maps effectively associate low-level features with the semantical motion patterns by exploiting the hidden information in crowded scenes. Motion pattern analyzing is implemented in a completely unsupervised way and the feature maps are clustered automatically through hierarchical clustering algorithm building on the basis of graphic model. The experiment results precisely reveal the distributions of motion patterns in current crowded videos and demonstrate the effectiveness of our approach.

Index Terms—crowded scene analysis, motion pattern, tracklet, the hybrid generative-discriminative feature maps, automatic clustering

1. INTRODUCTION

Crowded scenes analyzing is recently attracting much more attention in computer vision field. It has been an important and hot research topic because of its valuable potential applications. As shown in Fig.1, a crowded scene may include hundreds even thousands of objects, such as crowd, fauna, vehicles and so on. Crowded scenes arise commonly in our daily life, such as supermarket, high way, public meetings, etc. Capturing crowd dynamic is becoming increasingly more important and meaningful to scientific research and public security.

With the increase of density and complexity of objects and scenes, clearly exploring the situations of crowded scene becomes more challenging. Current techniques of visual surveillance on crowded level are still immature. In summary, the main difficulty in the research about crowded scenes is that the effective features from single object are very hard to extract because of its small size, low resolution, severe occlusions and similar appearance in diverse crowded scene. To overcome the difficulty, researchers are exploring new techniques according to the specific properties of crowded scenes.

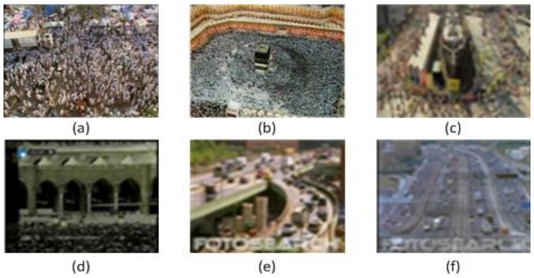

Fig. 1 Examples of crowded scenes.

T. Zhao et al. [1] perform human tracking in crowded scenes by modeling the human shape and appearance as articulated ellipsoids and color histograms respectively. This approach is one of the algorithms firstly applied in tracking in crowded scenes. The methods proposed in [2, 3] require that the scenes where objects are not moving densely and the tracking results of objects are available. However, these approaches are in absence of ability in dealing with high density crowded scenes.

To conquer the limitations of traditional methods, researchers are trying to study new methods based on the specialty of crowd scenes. Ali et al. [4] segment coherent crowd flows in video segmentation based on Lagrange Particle Dynamics. Ali et al. [5, 6] also track pedestrians by analyzing floor fields that describe how a pedestrian move based on scene-wide constraints. An underlying statistical model based on quantized SIFT features is used to capture the crowd properties in [7]. The steakline representation and potential functions in fluid dynamics are discussed to illustrate the crowd movement and detect motion patterns in [8]. X. Wang et al. [9] propose an unsupervised learning framework to model activities and interactions in crowded traffic scenes. C. Wang et al. [10] extract motion features based on Motion History Image. Bolei Zhou et al. [11] propose a new mixture model of Dynamic Pedestrian-Agents to learn the collective behavior patterns of the pedestrians.

In above related works, motion pattern detection and analysis play important and basic role. Under the context of crowded analysis, motion pattern refers to a spatial segmentation of the scene, within which a high degree of

978-1-4799-2341-0/13/$31.00 ©2013 IEEE

local similarity in speed as well as flow direction exist but otherwise outside [12]. Motion patterns not only describe the segmentation in the spatial space, but also reflect the motion tendency in a period. It can present the tendency of the crowd motion at a semantic level.

In this paper, we propose a novel approach to analyze motion patterns in dynamical crowded scenes based on the hybrid generative-discriminative feature maps. We make the following contributions: 1) Tracklets are collected by tracking dense feature points under three rules from the video of crowded scenes. 2) The hybrid generative-discriminative feature maps using posterior divergence are learned to capture the hidden moving information efficiently. 3) Dominant motion patterns are analyzed by hierarchical automatic clustering approach. We conduct experiment evaluations on the proposed approach and achieve satisfactory performance.

The rest of our paper is organized as follows. In section 2, 3 and 4, the details of our approach are discussed. The section 5 presents the experimental evaluations on diverse crowded videos. Finally in section 6, we make the conclusions.

2. DENSE POINTS TRACKING

Accurate and dense tracking from videos is an important requirement for crowded video surveillance. The most popular point tracker is Kanade-Lacas-Tomasi Tracker (KLT) [13]. However, KLT tracker seems to be unsuitable for most crowded scenes. Only a few points could be tracked due to the obscure structural features and severe occlusions in crowded scenes.

In crowded scenes, it's hard to track an individual for a long period because the inter- and intra-object occlusions make the problem suffering from tracking drift. The tracked point may be merged into another motion pattern at next frame or the optical flow vectors may not be estimated correctly, especially at the boundary of the motion. In this case, it’s very necessary to terminate the tracking when ambiguities caused by occlusion and scene mass arise.

Inspired by [14, 15, 16, 17], we propose three constraints to track the dense points according to the specific properties of crowded scenes. We collect tracklets by tracking densely sampled points using optical flow fields. A tracklet is a fragment of a trajectory during a short period.

We initialize the points at every pixel in the beginning as the flow field is dense. The optical flow field $w_t = (u_t, v_t)$ is calculated by the classical LK optical flow algorithm. Tracklets are obtained under the following three rules:

$$(x_{t+1}, y_{t+1}) = (x_t, y_t) + (M * w_t)\ |_{(\bar{x}_t, \bar{y}_t)} \tag{1}$$

$$|w + \hat{w}|^2 < \alpha_1 (|w|^2 + |\hat{w}|^2) + \beta_1 \tag{2}$$

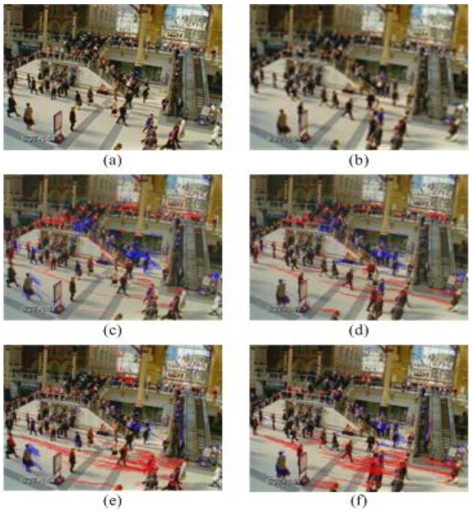

Fig. 2 Tracklets in the video of airport scene.
(a)(b) Original images. (c)(d) Trajectories by KLT tracker.
(e)(f) Tracklets by our method.

$$|\nabla u|^2 + |\nabla v|^2 > \alpha_2 |w|^2 + \beta_2 \tag{3}$$

In equation (1), M denotes the median kernel, $(\bar{x}_t, \bar{y}_t)$ is the rounded position of point (x_t, y_t) .Each point at frame *t* is tracked forward to the frame *t+1* by median filtering in a dense optical flow field. Global smoothness constraints *M* are employed among the points propagation, which is more robust than bilinear interpolation in [14], especially for points near motion boundaries.

In equation (2), $\hat{w} = (\hat{u}, \hat{v})$ denotes the flow from frame *t+1* back to *t*. In a non-occlusion situation, the backwrad flow vector should keep in accordance with the inverse direction of the forward flow vectors. A tolerance interval $\alpha_1(|w|^2 + |\hat{w}|^2) + \beta_1$ is allowed to increase linearly with the magnitude of flow vector.

In equation (3), ∇u and ∇v describe the divergence of the optical flow vector. In a non-occlusion situation, the divergence at one point should be similar to that of the surrounding. The third rule is proposed to stop the tracking to avoid a point drifts to the side of another motion boundary and mixes into different motions.

The tracklets with the three rules above have better quality than KLT tracker. The length of the tracklets is always short, but more reliable to reflect the ground truth of the crowd motion. This helps to improve the performance of seeking more effective motion information. Fig.2 shows the tracklets in crowded airport. Obviously, the trajectories obtained by KLT tracker in Fig.2(c)-(d) are sparse, and the noise caused by tiny disturbance accumulated gradually at the top of the image. Benefitting from our rules, the dense points can be tracked robustly, and the noise is greatly reduced.

3. HYBRID GENERATIVE-DISCRIMINATIVE FEATURE MAPS

The tracklets discussed above are variable-length sequences, and the points on tracklets are four-dimensional vectors including the location and velocity in horizontal and vertical directions. These low-level features are insufficient to describe samples directly.

Motion pattern presents the tendency of the crowd motion at a semantic level. So high-level semantic information based on tracklets should be derived to present motion patterns. We employ Posterior Divergence to mine the high-level semantic information.

Posterior Divergence [18, 19] is a hybrid scheme deriving feature mapping. It is an effective algorithm to derive explicit feature mapping from generative models, which captures discriminative information from samples by fully exploiting hidden variables.

For each tracklet, we train a HMM. The vectors on a tracklet are quantized into binary vectors according to the moving direction of each point. And then each tracklet is translated into a set of binary vectors.

Let $q^t_{1\times M}$ and $y^t_{1\times N}$ denote the hidden state (M probable state) and output state (N probable state), parameters π, $A_{M\times M}$, $B_{M\times N}$ be the initial state probability, transition probability and output probability. The model parameters $\theta = (\pi, A, B)$ can be learned by the Baum-Welch algorithm [20]. The free energy function of HMM is expressed as following:

$$F(Q,\theta) = E_Q[\sum_{i=1}^{M} q_i^0 \log\frac{\tau_i}{\pi_i} + \sum_{t=0}^{T_c}\sum_{i,j=1}^{M} q_i^t q_j^{t+1} \log\frac{g_{ij}}{a_{ij}} - \sum_{t=0}^{T_c}\sum_{i,j=1}^{M} q_i^t y_j^t \log b_{i,j}] \tag{4}$$

The approximate distributions $\{Q^i(q_i^0, q_i q_j, q_i y_j | \tau, g)\}$ is estimated based on the model θ, $q_i q_j$ and $q_i y_j$ can be viewed as two sets to variables. The feature map ϕ^c is derived as following:

$$\phi^c = \mathrm{vec}(\{\sum_{k\neq c} q_{ij}^k \log\frac{a_{i,j,+c}}{a_{ij}}, g_{ij}^c \log a_{i,j,+c}, g_{ij}^c \log g_{ij}^c\}_{i,j}^{M,M} \cup \{\sum_{k\neq c} h_{ij}^k \log\frac{b_{i,j,+c}}{b_{ij}}, h_{ij}^c \log b_{ij}\}_{i,j}^{M,M}) \tag{5}$$

The features maps are not visual features in the scenes, but are abstract ones with the fixed length. The feature maps effectively associate low-level features with the semantical motion patterns by exploiting the hidden information in crowded scenes. The feature maps both ensure the integrity of the features, and fully include the discriminating capacity. And also they can be straightforwardly worked for clustering scheme in analyzing phase efficiently. This greatly improves the performance of motion pattern analyzing.

4. MOTION PATTERN ANALYZING

According to the Gestalt theory of human visual perception, the main factors used in grouping are proximity, similarity, closure, simplicity and common fate [21]. According to this definition, "Motion pattern" in our research means the spatial-temporal tendency discrimination in a video. In this section, we cluster the feature maps with implicit motion information to analyze the motion patterns.

A novel hierarchical clustering method base on graphic model techniques is extended to our research. The algorithm makes clustering automatically without pre-defined number of clusters, and is effective for large size of data set [22].

This algorithm considers the cluster problem as a node-labeling processing on graphs. The procedure of clustering is constructed as the authority seeking on graph. The authority-shift procedure as following is performed iteratively for each Personalized Page Rank score (PPR) propagation until the n order PPR converges.

$$PPR_{n+1}(i) = PPR(PPR_n(i)) \tag{6}$$

$$Auth_n(i) = \arg\max PPR_n(i) \tag{7}$$

In our paper, a feature map is constructed as a vertex. The links between vertices reflect the relationship between feature maps. The weight reflects the similarity between the feature maps. This process is similar to mode-seeking in the mean-shift algorithm. But the difference is that the authority-shift only needs to be formulated once per node without special stopping rule.

5. EXPERIMENT

To test our approach on analyzing motion patterns in dynamic crowded scenes, we conduct experiments on typical videos in UCF_Crowds dataset, which are usually used to evaluate the performance of the algorithm about crowded scenes analysis. And we also implement experiments in our SJTU_Crowds dataset. We compare our approach with the state-of-the-art approach in [8].

The UCF_Crowds dataset is constructed by the University of Central Florida. The dataset contains videos of crowds， high density of moving vehicles, bio-cells under microscopes.

The SJTU_Crowds dataset is designed to facilitate the research about crowd analysis. We collect the dataset in a square of Shanghai Jiao Tong University campus. This dataset is different from the UCF_Crowds dataset in densities, motions, scenes, and so on. We will public this SJTU_Crowds dataset later.

The experimental results are shown in Fig.3 and Fig.4. A video in a large market with complex and high density crowd is described in Fig.3 (a). Thousands of people are moving in the market. The individuals can't be discriminated due to the similar appearance, extremely occlusion with small size. In this case, it's difficult and also unnecessary to

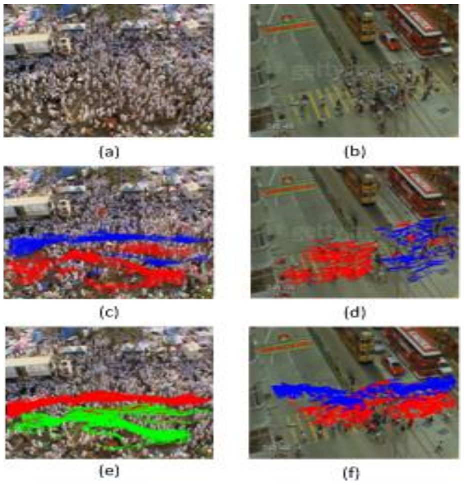

Fig.3 The results of videos. (a) Market. (b) Crossroad. (c)(d) Tracklets. (e)(f) Motion patterns.

distinguish individuals. The tracklets shown in Fig.3 (c) capture the persistent motion information. Two dominant motion patterns are detected marked in different color in Fig.3 (e). Another scene with the individuals crossing at the crossroad is shown in Fig.3 (b). The individuals are tracked in Fig.3 (d), and the two dominant motion patterns with opposite directions are detected successfully.

The videos of two and five queues are shown in Fig. 4. As can be seen from the results, the tracklets reflect the reliable movement information of the queues in Fig.4 (c) and (d). It seems difficult to detect motion patterns while individuals in different queues move close to others. However, the motion patterns are detected precisely in Fig.4 (e) and (f) by our method.

In order to evaluate the performance of our algorithm, we select six typical videos on both datasets. We compare our results with the ground truth and the state-of-the-art method. The ground truth is manually marked from the videos. The comparisons about the number of dominant motion pattern detected by different methods are shown in Fig.5 and Fig.6. From the comparisons, we can see that the results of our method are accordant with the ground truth in most cases. The advantages of our method are significant. One major reason is that the tracklets can extract the valuable motion motion information, and then the feature maps also reflect the reliable and discriminative implicit motion information

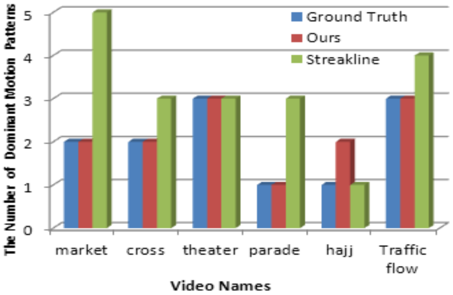

Fig.5 The performance evaluation on UCF_Crowds dataset.

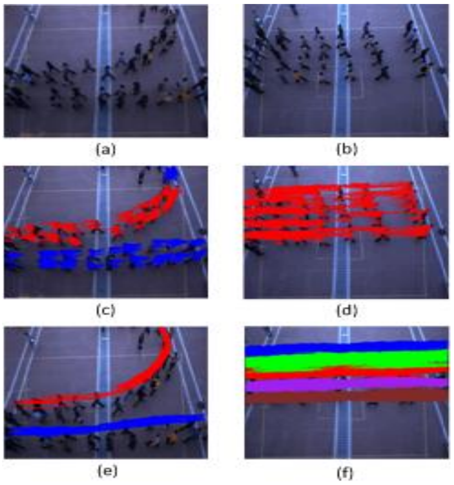

Fig.4 The results of videos. (a) Two queues. (b) Five queues. (c)(d) Tracklets. (e)(f) Motion patterns.

with fixed length. And the feature maps can be straightforwardly worked for clustering scheme in analyzing phase efficiently.

6. CONCLUSION

In this paper, we propose an unsupervised method to analyze motion patterns in dynamic crowded scenes. We track dense points under three rules through LK optical flow algorithm. And then the motion patterns are analyzed by the automatic hierarchical clustering algorithm with feature maps derived by hybrid generative-discriminative scheme. The experiments are conducted on some challenging videos in UCF_Crowds dataset and our SJTU_Crowds dataset. The results precisely reveal the distributions of motion patterns in current crowded videos and demonstrate the effectiveness of our approach.

We plan to investigate more effective features about the crowd, and further use this research for scene understanding in crowded scenes.

7. ACKNOWLEDGEMENTS

This work was supported in part by National Basic Research Program of China (973 Program) 2011CB302203 and the National Science Foundation of China under Grant 61273285.

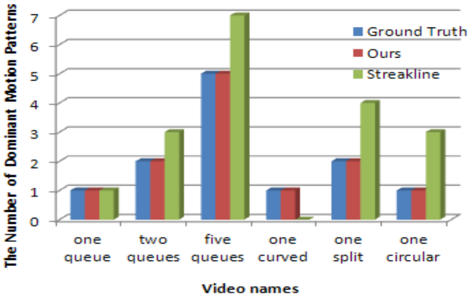

Fig.6 The performance evaluation on SJTU_Crowds dataset.

8. REFERENCES

[1] Zhao, T. and Nevatia, R., “Tracking multiple humans in crowded environment,” Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004, Princeton, NJ, USA., Vol. 2, pp. II–406.

[2] Khan, Z. and Balch, T. and Dellaert, F., “An MCMC-based particle filter for tracking multiple interacting targets,” Computer Vision-ECCV 2004. Lecture Notes in Computer Science Volume 3024, 2004, pp. 279-290

[3] Sugimura, D. and Kitani, K.M. and Okabe, T. and Sato, Y. and Sugimoto, A., “Using Individuality to Track Individuals: Clustering Individual Trajectories in Crowds Using Local Appearance and Frequency Trait,” Proc. of IEEE Int’l Conf on Computer Vision, 2009.

[4] Ali, S. and Shah, M., “A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis,” 2007 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2007, pp. 1–6.

[5] Ali, S. and Shah, M., “Floor fields for tracking in high density crowd scenes,” European conference on computer vision, Vol. 2, 2008, pp. 1–14.

[6] Ali, S., Taming crowded visual scenes, Ph.D. thesis, University of Central Florida (2010).

[7] Arandjelovic, O. and Zisserman, A., “Crowd detection from still images,” British Machine Vision Conference, Vol. 1, 2008, pp. 523-532.

[8] Mehran, R. and Moore, B. and Shah, M., “A streakline representation of flow in crowded scenes,” Computer Vision–ECCV 2010 (2010), 439–452.

[9] Wang, X. and Ma, X. and Grimson, W.E.L., “Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models,” Pattern Analysis and Machine Intelligence, IEEE Transactions on 31 (3) (2009), pp. 539–555.

[10] Wang, C. Zhao, X. and Zou, Y. and Liu, Y., “Detecting motion patterns in dynamic crowd scenes,” 2011 Sixth International Conference on Image and Graphics (ICIG), IEEE, 2011, pp. 434–439.

[11] Zhou, B. and Wang, X. and Tang, X., “Understanding collective crowd behaviors: Learning a Mixture model of Dynamic pedestrian-Agents,” 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012, 2871–2878.

[12] Saleemi, I. and Hartung, L. and Shah, M., “Scene understanding by statistical modeling of motion patterns,” 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2010, pp. 2069–2076.

[13] Shi, J. and Tomasi, C., “Good features to track,” IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1994. IEEE, 1994, pp. 593–600.

[14] Sundaram, N. and Brox, T. and Keutzer, K., “Dense point trajectories by GPU-accelerated large displacement optical flow,” 2010 IEEE European conference on computer vision (ECCV). IEEE, 2010, pp. 438–451.

[15] Wang, H. and Klaser, A. and Schmid, C. and Liu, C.L., “Action recognition by dense trajectories,” 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2011, pp. 3169–3176.

[16] Martin, R. and Arandjelovic, O., “Multiple-object tracking in cluttered and crowded public spaces,” Advances in Visual Computing, 2010, pp. 89–98.

[17] Arandjelovic, O., “Contextually Learnt Detection of Unusual Motion-Based Behaviour in Crowded Public Spaces,” Computer and Information Sciences II, 2012, pp. 403–410.

[18] Li, X. and Lee, T.S. and Liu, Y., “Hybrid generative-discriminative classification using posterior divergence,” 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011, pp. 2713–2720.

[19] Wu, Z. and Li, X. and Zhao, X. and Liu, Y., “Hybrid generative-discriminative recognition of human action in 3D joint space,” Proceedings of the 20th ACM international conference on Multimedia, 2012, pp. 1081--1084.

[20] Rabiner, L.R., “A tutorial on hidden Markov models and selected applications in speech recognition,” Proceedings of the IEEE, 77(2):257-286, 1989

[21] Hu, M. and Ali, S. and Shah, M., “Learning motion patterns in crowded scenes using motion flow field,” Proceedings of the 19th International Conference on Pattern Recognition (ICPR’08), Citeseer, 2008.

[22] Cho, M. and MuLee, K., “Authority-shift clustering: Hierarchical clustering by authority seeking on graphs,” 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2010, pp. 3193–3200.