SUPER-PIXEL BASED CROWD FLOW SEGMENTATION IN H.264 COMPRESSED VIDEOS

Sovan Biswas *R. Gnana Praveen* *R. Venkatesh Babu*

Video Analytics Laboratory
Supercomputer Education and Research Centre
Indian Institute of Science
Bangalore, India

ABSTRACT

In this paper, we have proposed a simple yet robust novel approach for segmentation of high density crowd flows based on super-pixels in H.264 compressed videos. The collective representation of the motion vectors of the compressed video sequence is transformed to color map and super-pixel segmentation is performed at various scales for clustering the coherent motion vectors. The number of dynamically meaningful flow segments is determined by measuring the confidence score of the accumulated multi-scale super-pixel boundaries. The final crowd flow segmentation is obtained from the edges that are consistent across all the super-pixel resolutions. Hence, our major contribution involves obtaining the flow segmentation by clustering the motion vectors and determination of number of flow segments using only motion super-pixels without any prior assumption of the number of flow segments. The proposed approach was benchmarked on standard crowd flow dataset. Experiments demonstrated better accuracy and speedup for the proposed approach compared to the state-of-the-art methods.

Index Terms— Crowd Flow Segmentation, Super-pixels, H.264 Compressed domain, Motion Segmentation

1. INTRODUCTION

With the increase in population, there is an imperative need to monitor and model large gatherings such as religious festivals, concerts, etc in order to avoid any catastrophic events such as stampedes, etc. In the recent years, analyzing and modeling the videos involving high density crowded scenes have drawn attention of several researchers from various perspectives, where computer vision algorithms have played a significant role to address this problem. However the performance of these traditional vision approaches gets deteriorated with the increase in density due to the complex dynamic behavior. Hence there is an immense need to address the problem of analyzing and modeling high density crowded scenes such as shown in Fig. 1.

Most of the existing approaches for the problem of crowd flow modeling was done in pixel domain. In our approach, we had explored the prospect of performing crowd flow segmentation in H.264 videos as it has found a wide range of applications due to its high resolution, low bandwidth usage, etc. As decoding the compressed videos is additional computation, we explored the possibility of reducing the computation through partial decoding of H.264 compression parameters and use them as features for crowd flow segmentation. This drastically reduces computation costs and simultaneously provide initial segmentation to built further pixel level processing.

Since the information available in the compressed domain is very limited compared to the pixel domain, it is highly challenging to address this problem in compressed domain. However, we had achieved crowd flow segmentation by utilizing only motion vectors as it conveys significant information related to the dynamics of crowd flow behavior.

Fig. 1. Examples Scenarios of videos involving high density crowded scenes in [1]

The subsequent portion of the paper is organized as follows. The related work for crowd flow analysis is discussed in Section. 2. Section. 3 formulates the problem of crowd flow segmentation and various challenges involved in it. The proposed approach was expounded in Section. 4. The results on the database provided by [1] is discussed and analyzed in Section. 5. Finally, the conclusion of the proposed approach and the scope for future work was outlined in Section. 6.

2. RELATED WORK

Over the past few years, several vision researchers have explored different problems related to crowd analysis from various perspectives. Most of the work done so far in crowd analysis is mainly focussed on crowd detection, tracking of individuals [2], measuring crowd collectiveness [3], detection of anomalous behavior [4, 5], etc. Jacques et al. [6] and Zhan et al. [7] provided a detailed review of various vision approaches in order to tackle different problems in crowd analysis. In this work, we have focussed on the problem of segmenting the dominant and dynamically meaningful flow segments in crowded scenarios involving high density moving objects.

Wu et al. [8] exploited translational flow to approximate local crowd motion and developed a region growing scheme for crowd flow segmentation based on optical flow field. They had also proposed another approach for crowd flow segmentation based on fuzzy c -means clustering [9]. Wu et al. [10] had proposed an approach for crowd flow partitioning by perceiving this problem as a problem of scattered motion field segmentation by assuming the local crowd motion as translational motion field. They had implemented their approach on real crowded sequences and shown better results in segmenting homogeneous regions in the crowded scenes. In another work by Li et al. [11], the dynamic region that corresponds to

the crowd flow is extracted and the flow segmentation is performed based on the histogram of the orientation of the foreground velocity field. Anil et al. [12] had outlined the review of various computer vision algorithms dealing with crowded scenarios and proposed a system for the automatic detection of dominant patterns of crowd flow in dense crowd scenarios by tracking the low level object features using the optical flow algorithm. Kuhn et al. [13] had developed a frame work for extracting the motion patterns by combining the optical flow with Lagrangian analysis of time dependent vector fields shown its applicability for crowd analysis like automated detection of abnormal events in the video sequences. Another significant work to address this problem was done by Ali et al. [1], where they have modeled this problem from the perspective of fluid dynamics. A grid of particles was superimposed on the video and trajectories of these particles over the frames of the video was obtained using optical flow and further processed to compute the final crowd flow segmentation. They had also extended their approach for the detection of instabilities such as anomalous behaviors in the crowded scenes.

All the above mentioned approaches have dealt this problem in pixel domain where optical flow is computed for the crowd video sequence and further processed to achieve the crowd flow segmentation. Since it is computationally expensive to obtain the optical flow, it may not be useful for real time applications. Hence we propose a novel framework of achieving flow segmentation in the compressed domain using motion vectors. The prospect of achieving crowd flow segmentation in the compressed domain was explored in [14]. However, on the pursuit of improving the computational efficiency, a novel framework with less computational complexity still retaining comparable accuracy was proposed and demonstrated on the dataset provided by [1].

3. PROBLEM FORMULATION

We have perceived the problem of crowd flow segmentation from the perspective of determining the number of physically and dynamically meaningful flow segments using flow vectors (F). In order to obtain the number of flow segments, the delineation between the dynamically meaningful segments have to be efficiently captured while discarding the spurious regions which do not contribute to the demarcation of semantic crowd flow segments. In the proposed approach, this is performed through multi-scale image representation obtained using super-pixel segmentation with varying number of segments.

Let E_k be the set of edges that corresponds to the scale k where, $1 \leq k \leq S$, S represents the number of scales (number of super-pixels) which is shown in eqn. (1)

$$E_k = \{x_i : \forall x_i \in \partial\mathcal{R}_k\} \tag{1}$$

where, x_i represents the pixels along the edges of the super-pixels at scale k, $\mathcal{R}_k$ denotes the segments and $\partial\mathcal{R}_k$ denotes the boundaries of the super-pixel segments.

Let E be the entire set of edges at all the scales which constitutes both weak (spurious edges) and strong edges as shown in eqn. (2).

$$E = \bigcup_{k=1}^{S} E_k \tag{2}$$

The detection of the significant edges helps in computing the final crowd flow segmentation. Now the problem has converged to computing the edge-set E_f which is the subset of E ($E_f \subset E$) where, the edge set E_f denotes the edges that distinguishes the significant crowd flow in the video sequence as shown in eqn. (3).

$$E_f = \{x_i : \forall x_i \in \partial\mathcal{R}_g\} \tag{3}$$

where, x_i represents the pixels of the edges of the final edge map,$\mathcal{R}_g$ denotes the ground truth segments and $\partial\mathcal{R}_g$ denotes the boundaries of the regions of the ground truth for crowd flow segmentation.

4. THE PROPOSED APPROACH

The proposed approach constitutes of four stages. The overall block diagram of the proposed approach is shown in Fig. 2.

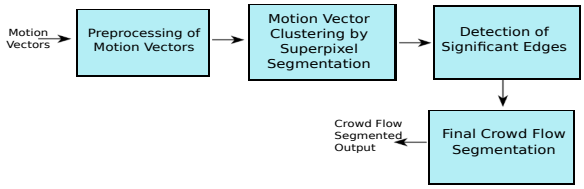


Fig. 2. Overview of the proposed approach

4.1. Preprocessing of Motion Vectors

Motion Vectors of H.264 compressed videos are aimed at video compression through exploring the temporal redundancy to optimize the number of encoding bits. As this optimization depends upon the underlying compression profile and algorithm used, they may not follow true object motion across consecutive frames and can be considered as noisy for the current purpose. So the noisy motion vectors which shoots out at the static region of the video has to be eliminated before further processing of motion vectors. It plays a pivotal role in computing the accurate crowd flow segmentation. Hence the motion vectors that contribute to the crowd flow is retained and erroneous motion vectors are discarded by taking spatio-temporal median filtering (with a neighborhood of $5 \times 5 \times 5$). Still there can be some erroneous motion vectors which is discarded by eliminating the motion vectors that occurs only for the few frames. Empirically, the motion vectors which occurs less than 10 percent of the number of frames at a specific location is discarded.

Secondly, there can be camera motion which leads to nonzero motion vectors for the static region resulting in distortion of the background. Hence the motion vectors in the background is eliminated by incorporating camera motion compensation process as mentioned in Biswas et al. [15] which is shown in eqn. (4)

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = s \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} p_3 \\ p_4 \end{pmatrix} \tag{4}$$

where, s is the scale factor, p_3 and p_3 are the pan rate and tilt rate respectively. (x, y) and (x', y') represents the current and future locations of the blocks. After discarding the erroneous motion vectors and compensating for the camera motion, a motion pattern is generated by considering the collective representation of the motion vectors, which is obtained by taking the median of the motion vectors temporally over all the frames. In other words, the resultant motion vectors at a particular location (i, j) is obtained by taking the median of all the motion vectors temporally at that particular location. Subsequently, a motion mask is generated corresponding to regions with motion information.

4.2. Clustering of Motion Vectors by Super-pixels

The resultant motion vectors obtained from the previous step are transformed to color space as mentioned in [16] and the color coded map is smoothed in order to avoid the blocky effect due to the

sub-macro block level motion compensation. Now the super-pixel segmentation is performed on this color coded map at various levels/scales by varying the number of super-pixels. Several researchers have explored various algorithms for super-pixel segmentation. However, in this work we have used the super-pixel segmentation proposed by Li et al. [17]. As we increase the number of super-pixels, it results in clustering the motion vectors from coarser to finer level and subsequent extraction of edges of clustered motion through super-pixel boundaries. But the strong edges that separates the distinctive flow patterns of the scene and the delineation that discriminates the orientation of the motion vectors is captured in most of the levels. Hence these strong edges which contribute to the delineation of the crowd flow segmentation is retained by exploiting the strength of these edges and orientation of the motion vectors on both sides of the corresponding edges. The intermediate stages of the proposed approach is shown in Fig. 3.

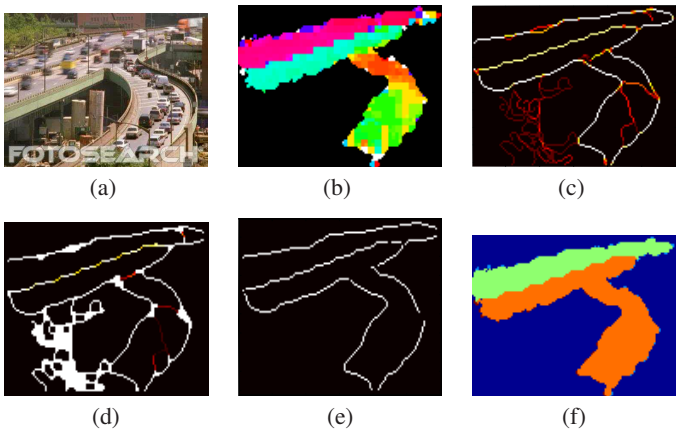


Fig. 3. Output at the intermediate stages of the proposed approach (a) Input Video Sequence (b) Color Coded Version (c) Confidence Score C_1 (d) Confidence Score C_2 (e) Final Confidence Score C_f (f) Flow Segmentation

4.3. Detection of true edges using Confidence Scores

The segmented output of boundaries of the super-pixels at various levels are integrated together and a collective representation of the segmented output was obtained which contains the edges at all the levels. We have introduced two confidence scores to retain the true edges of the super-pixels and prune the spurious edges at various levels. The first confidence score is based on the strength of the edges. This is obtained by integrating all the corresponding edges at various levels of the super-pixels as shown in eqn. (5)

$$\mathcal{E}_s(x_i) = \sum_{k=1}^{S} E_k(x_i) \tag{5}$$

where, $\mathcal{E}_s$ denotes the sum of edge strength across all the levels, E_k represents the boundaries of super-pixels at level k, S denotes the number of scales and x_i is the pixel values of the boundaries of the regions.

Since the true edges which contributes to the crowd flow segmentation is retained in most of the levels, it will have higher strength compared to the spurious edges which occurs only in a few levels. The first confidence score (C_1) is obtained by normalizing the resultant integrated edge map as shown in eqn. (6)

$$C_1 = \frac{\mathcal{E}_s}{S} \tag{6}$$

where, $\mathcal{E}_s$ is the edge strength for all edges as defined in eqn. (5). S is the number of scales.

Secondly, the second confidence score (C_2) is measured based on the orientation of the motion vectors on both sides of the edges. The difference of orientation of the motion vectors on both sides of the edges is obtained to discard the edges that contribute to over segmentation which is shown in eqn. (7)

$$C_2 = \frac{D(\angle\theta_{S_i} - \angle\theta_{S_j})}{\pi}, \;\; s.t. \;\; S_i \in N(S_j) \tag{7}$$

where, $\angle\theta_{S_i}$, $\angle\theta_{S_j}$ represents the orientation on both sides of the corresponding edge and $D(\angle\theta_{S_i} - \angle\theta_{S_j})$ represents the angular distance between the orientations on both sides of the edges. $N(S_j)$ denotes neighbors of S_j. If the orientation on both sides of the motion vectors are coherent to each other, then the difference of orientation of the corresponding edge will be very low and vice-versa. Now the final confidence score for each edge is obtained by multiplying the two confidence scores as shown in eqn.(8)

$$C_f = C_1 \times C_2 \tag{8}$$

4.4. Final Crowd Flow Segmentation

Now the final confidence score from the previous step is refined by the motion mask obtained during preprocessing (see Sec. 4.1) to suppress any spurious edges emerging in the static region of the scene. The resultant final score is then thresholded to R_τ, where R_τ represents the set of edges whose strength is more than τ as shown in eqn. (9)

$$R_\tau(x, y) = \begin{cases} 0 & \text{if } C_f < \tau \\ 1 & \text{otherwise} \end{cases} \tag{9}$$

where, τ is varied from 0 to 1 to extract the corresponding edge set R_τ.

Now the true edges that corresponds to the contribution of crowd flow segmentation (E_f) is obtained by following the steps given in the algorithm 1.

Algorithm 1

Require: Threshold $(\mathcal{T}) = \{\tau_i\} \;:\; 1 \leq i \leq t$ and corresponding R_{τ_i}

Ensure: E_f as defined in Eq. 3

for i = 1 to t **do**

$J_i = \max_{E_k} J(R_{\tau_i}, E_k)$ where, J_i is the maximum Jaccard measure between R_{τ_i} and $E_k \; \forall \; 1 \leq k \leq S$

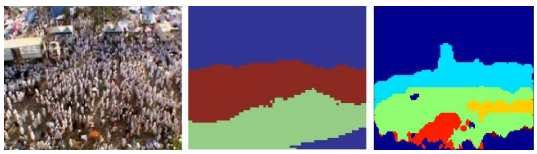
$E^*_{k_i} = \operatorname{argmax}_{E_k} J(R_{\tau_i}, E_k)$

end for

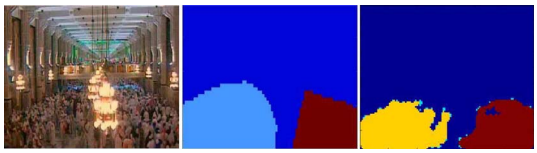
$E_f = \operatorname{argmax}_{E^*_{k_i}} J_i$

(a) Sequence 2

(b) Sequence 3



(c) Sequence 4



(d) Sequence 6

Fig. 4. Experimental Results of some of the videos in the database. First column of images shows input video sequence, second column of images show Ali et al.[1], whereas third column shows the output of the proposed approach

5. RESULTS AND DISCUSSIONS

5.1. Experiments

The proposed approach is evaluated on the dataset provided by Ali et al. [1], which contains a wide range of dynamics. Since, the dataset was not encoded in H.264 format, we have first encoded the same in H.264 format using x264[1] baseline profile (only I and B frames) with 1 reference frame and Group of Pictures (GOP) length is set to 30. Since B frames are not used, baseline profile is ideal for network cameras and video encoders to achieve low latency [18]. The performance comparison of the proposed approach with Ali et al. [1] is bench-marked using Jaccard measure with the manually generated groundtruth. The motion vectors are extracted by partially decoding each H.264 video sequence and preprocessed as described in section. 4.1. These extracted motion vectors varies from sub-macro block of 4×4 to 16×16, but for ease of computation we replicate the motion for each macroblocks to its constituent 4×4 blocks. The number of super-pixels are varied in the range of 3 to 10 for all the experiments. Few of sample crowd flow segmentation obtained through algorithm 1 are shown in Fig. 4. All the experiments were performed using MATLAB on single core 3.4 GHz processor.

5.2. Analysis

Since we are using only motion vectors of H.264 compressed videos, the proposed approach runs faster for this challenging dataset. Some of the qualitative results of flow segmentation for the videos in the dataset are shown in Fig. 4. The Jaccard measure obtained for the video sequences are shown in table 1 (in the same order as that of the Fig. 4). Since Ali et al. [1]'s method is a global approach, the output flow segments extend beyond its actual demarcations, whereas our proposed approach appropriately captures the demarcations of the dynamic regions of the flow. Hence the proposed approach, even with noisy motion vectors shows better accuracy than that of [1] as shown in Table. 1. For instance, the traffic video sequence (Seq 2) shown in Fig. 4 clearly shows that the proposed approach captures the dynamic flow of the crowded scene with better precision than that of the [1]. The Sequence 4 in the dataset is a complex flow pattern where people are moving alternatively in random directions. Since Ali et al.'s[1] approach performs segmentation at the global level, it merges the diverse pattern and gives only the dominant flows, whereas the proposed approach could capture the diverse variation of the complex flow pattern as shown in Fig 4.

It is observed in table1, the performance degrades for Sequence 5 and 7. This is due to the fact the motion vectors in these videos could not capture very fine motion information at sub-pixel accuracy. Sequence 5 is of frame size 188×144 compared to 480×360 for rest of videos. For Sequence 7 (mecca sequence), most of the motion vectors are not retrieved due to the adjacent noisy motion vectors. Hence as long as the motion information is captured by motion vectors, the proposed approach performs better or equivalent to [1].

The execution time for the proposed approach is approximately 5 secs for a video with 100 frames of size 480×360 with unoptimized matlab code compared to 30 sec by Ali et al. [1], without optical flow computation, executed on the same machine.

Table 1. Jaccard Index Similarity Measure for [1] and our proposed approach for some of the videos in the database with reference to ground truth

Video Sequences	Jaccard Similarity Measure		
	Ali et al [1]	Proposed	Timings (secs)
Sequence 1	**0.63**	0.60	4.96
Sequence 2	0.28	**0.67**	5.08
Sequence 3	0.57	**0.74**	4.66
Sequence 4	0.67	**0.68**	4.49
Sequence 5	**0.78**	0.24	4.32
Sequence 6	0.41	**0.62**	5.32
Sequence 7	**0.54**	0.15	4.95

6. CONCLUSION AND FUTURE WORK

We have proposed a novel approach for crowd flow segmentation by using only motion vectors in the compressed domain. The preprocessed motion vectors are clustered using super-pixel segmentation at various levels by varying the number of super-pixels, and the crowd flow segmentation was achieved by retaining the strong motion boundaries and discarding the spurious ones. The proposed approach was found to perform faster with better accuracy than pixel domain approaches. Since the proposed approach relies on the motion vectors, the performance of the proposed approach degrades for video sequences where the motion vectors of the dynamic region of the crowd flow is not properly captured.

7. ACKNOWLEDGMENT

This work was supported by CARS (CARS-25) project from CAIR, DRDO, Govt. of India.

[1] available at : http://www.videolan.org/developers/x264.html

8. REFERENCES

[1] S. Ali and M. Shah, "A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1–6.

[2] M. Rodriguez, J. Sivic, I. Laptev, and J.-Y. Audibert, "Data-driven crowd analysis in videos," in *Proceedings of the IEEE International Conference on Computer Vision*, 2011, pp. 1235–1242.

[3] Bolei Zhou, Xiaoou Tang, and Xiaogang Wang, "Measuring crowd collectiveness," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3049–3056.

[4] V. Mahadevan, Weixin Li, V. Bhalodia, and N. Vasconcelos, "Anomaly detection in crowded scenes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 1975–1981.

[5] Sovan Biswas and R. Venkatesh Babu, "Real-time anomaly detection in H.264 compressed videos," in *Proceedings of the National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics*, 2013.

[6] J.C.S Jacques Junior, S. Raupp Musse, and C.R. Jung, "Crowd analysis using computer vision techniques," *IEEE Signal Processing Magazine*, vol. 27, no. 5, pp. 66–77, 2010.

[7] Beibei Zhan, DorothyN. Monekosso, Paolo Remagnino, SergioA. Velastin, and Li-Qun Xu, "Crowd analysis: a survey," *Machine Vision and Applications*, vol. 19, no. 5–6, pp. 345–357, 2008.

[8] Si Wu, Zhiwen Yu, and Hau-San Wong, "Crowd flow segmentation using a novel region growing scheme," in *Proceedings of the Advances in Multimedia Information Processing*, 2009, vol. 5879, pp. 898–907.

[9] Si Wu, Zhiwen Yu, and Hau-San Wong, "A shape derivative based approach for crowd flow segmentation," in *Proceedings of the Asian conference on Computer Vision*, 2009, vol. 5994, pp. 93–102.

[10] Si Wu and Hau San Wong, "Crowd motion partitioning in a scattered motion field," *IEEE Transactions on Systems, Man, and Cybernatics, Part B: Cybernatics*, vol. 42, no. 5, pp. 1443–1454, 2012.

[11] Wei Li, Jiu-Hong Ruan, and Hua-An Zhao, "Crowd movement segmentation using velocity field histogram curve," in *Proceedings of the International Conference on Wavelet Analysis and Pattern Recognition*, 2012, pp. 191–195.

[12] A.M. Cheriyadat and R.J. Radke, "Detecting dominant motions in dense crowds," *IEEE Journal of Selected Topics in Signal Processing*, vol. 2, no. 4, pp. 568–581, 2008.

[13] A. Kuhn, T. Senst, I. Keller, T. Sikora, and H. Theisel, "A lagrangian frame work for video analytics," in *Proceedings of the IEEE International Workshop on Multimedia Signal Processing*, 2012, pp. 387–392.

[14] R. Gnana Praveen and R.Venkatesh Babu, "Crowd flow segmentation based on motion vectors in H.264 compressed domain," in *Proceedings of the IEEE International Conference on Electronics, Computing and Communication Technologies*, 2014, pp. 1–5.

[15] Sovan Biswas and R.Venkatesh Babu, "H.264 compressed video classification using histogram of oriented motion vectors (HOMV)," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 2040–2044.

[16] Simon Baker, Daniel Scharstein, J. P. Lewis, Stefan Roth, Michael J. Black, and Richard Szeliski, "A database and evaluation methodology for optical flow," *International Journal of Computer Vision*, vol. 92, no. 1, pp. 1–31, 2011.

[17] Ming-Yu Liu, Oncel Tuzel, Srikumar Ramalingam, and Rama Chellappa, "Entropy rate superpixel segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 2097–2104.

[18] "`http://www.axis.com/products/video/about_networkvideo/compression_formats.htm`,".