# Detection of Groups of People in Surveillance Videos based on Spatio-Temporal Clues

Rensso V. H. Mora Colque[†], Guillermo Cámara-Chávez[‡], William Robson Schwartz[†]

[†]Computer Science Department, Universidade Federal de Minas Gerais
Belo Horizonte, MG, Brazil
soulchicano@gmail.com, william@dcc.ufmg.br
[‡]Computer Science Department, Federal University of Ouro Preto
Ouro Preto, MG, Brazil
gcamarac@gmail.com

**Abstract.** Video surveillance has been widely employed in our society in the past years. In this context, humans play an important role and are the major players since they are responsible for changing the state of the scene through actions and activities. Therefore, the design of automatic methods to understand human behavior and recognize activities are important to determine which subjects are involved in an activity of interest. The computer vision research area has contributed vastly for the development of methods related to detection, tracking and recognition of humans. However, there is still a lack of methods able to recognize higher level activities (e.g., interaction among people that might be involved in an illegal activity). The first step to be successful in this enterprise is to detect and locate groups of people in the scene, which is essential to make inferences regarding interactions among persons. Aiming at such direction, this paper presents a group detection approach that combines motion and spatial information with low-level descriptors to be robust to situations such as partial occlusions. The experimental results obtained using the PETS 2009 and the BEHAVE datasets demonstrate that the proposed combination indeed achieves higher accuracies, indicating a promising direction for future research.

**Keywords:** Group detection, group activity, low-level descriptors, collective behavior.

## 1 Introduction

Humans are the main focus in the monitoring since they are the agents responsible for performing actions that change states of the scene. For instance, a human might interact with objects in the scene to accomplish some goal, such as the removal of an object from the trunk of a vehicle, or interact with other people to accomplish something that may be characterized as a suspicious activity. Therefore, the design of processing methods focusing on humans is extremely important to being able to determine what is the role of each person in the

scene so that responsibilities can be set to each agent, such as to determine which subjects are involved in a specific activity.

Since interactions among humans provide relevant information for activity recognition and understanding, the analysis of images and videos involving humans, presents large interest of the research community. In this scope, solving Computer Vision problems such as feature extraction, background subtraction, pedestrian detection, face recognition, person tracking, person re-identification, gesture recognition, pose estimation, action recognition, and activity recognition [1] are fundamental to model interactions among agents aiming at understanding high-level activities performed in a scene.

In this paper, we present an approach to detect groups of people in video sequences. The method is based on three main steps: segmentation, tracking and grouping. One of the main advantages of the model is its simplicity to group people in the videos, because it uses basic grouping rules: proximity combined with direction and local information. The first rule takes in count the trajectory of moving segments, and second is a clue given by a classifier that has been trained with spatio-temporal features based in a local descriptor. Separately, these two rules do not provide accurate results, but their combination improves the results due to the employment of the complementary sources of information.

The experimental evaluation considers situations in which both small and large number of people are in the scene. The experiments, performed using the PETS 2009 and the BEHAVE [2] datasets demonstrate the viability of the method to detect groups of people in an accurate manner.

## 2 Related Works

The literature shows that exits a great interest in activity recognition focusing both on individuals or groups. In [3–5] demonstrate the wide variety of research on the recognition of human activities with the proposition of several taxonomies to classify approaches that have been proposed for activity recognition. Particularly, Vishwakarma and Agrawal [5] presented an interesting general flow line to recognize a human activity: *motion segmentation, object classification, human tracking, action recognition and semantic description*, common steps in the majority of the approaches in the literature. In addition, they presented an overview of the various types of human activities as a function of the complexity level and time span, in which higher levels imply recognition of previous levels.

Now, we will describe specific models for group activity recognition. A model widely used is based on the description of the trajectories of people to determine a grouping criterion. Hongeng and Nevatia [6] proposed a method where a single thread of action is recognized from the characteristic of the trajectory. Similarly, Ryo and Aggarwal [7] proposed a set of complex rules based on trajectories to define the interactions of individuals within a group, where the extracted information of the scene is processed using a classifier and finally recognizing one of the following situations: group activity, people-group interaction, group-group interaction, inter-group interaction, also includes combinations of previous

situations. In addition, Ni and Kassim [8] divide the activities in individual, pair and group situations. Recognition is finally given by a study of trajectories and a set of rules for different situations. Adaptive Mean Shift is used to track the individuals and the Gaussian process is used to cluster people in groups [9].

Other types of approaches perform comprehensive analysis of the trajectories [10], mixing background removal and optical flow to extract the global movement of each frame image. The grouping criteria is aided by a classifier to determine the action. Motion models could be predicted by Hidden Markov Models [11], k-means algorithm [12] or multi-agents for semantic comprehension [13]. For instance, Chang *et al.* [14] perform the recognition based on the information of the trajectories of the actors capture by multiple cameras. Histograms of velocities are also used to understand the crowd behavior [15].

## 3 Proposed Approach

To incorporate both, tracking and low-level descriptors, the proposed method relies in the following steps: moving object segmentation, tracking segment, flow estimation and local feature description. In first step, the method segments people using the motion in the sequence captured by background subtraction obtained with Mixture of Gaussian (MOG) model [16]. Once the segments have been obtained, we use optical flow to determine the flow of moving segments, this step allows to track the segments in the sequence. The next step consists in obtain the clues that allow us to group people. First, flow orientation histograms are builded to determine the direction of an entire segment, thus, segments that are moving in same direction may be considered as group in some specific time. The other clue is given by the result of the segment classification, in this part we made an adaption to the Histogram of Gradients(HoG), adding time information to the descriptor. The steps of our approach are illustrated in Figure 1.

**Segmentation and Tracking.** Given an input video sequence, the first step of the process is to determine the background image employing the Mixture of Gaussians (MoG) technique [16] to generate progressively the background image. After that, the moving segments are extracted using frame difference. Basic image processing techniques, such as median filter and morphological operations, are employed to eliminate noise. Segments that correspond to human bodies are distinguished from other objects according to their size.

Once the segments are defined, the optical flow is estimated [17] for them. Tracking segments is simple, since optical flow predicts the position in next frame, the corresponding segment is situated according to a proximity criterion using the prediction calculated in previous frame.

**Grouping: Flow.** One of the challenges on group detection is to get the grouping rules. In our case, the segments obtained in the previous step might contain one or more people. An initial grouping criterion consists of two main rules, join the segments with same direction, which means walking or running in the same direction, and join them if they are close enough. The minimum distance that

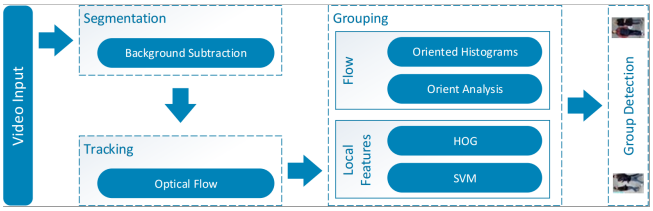


Fig. 1: Our proposed approach. The method is basically divided in three major steps. In first, a segmentation process extracts moving objects from the scene using background subtraction based on the mixture of Gaussians [16]. Then, the segments are tracked by optical flow. In the final step, the relationships between segments and their labels are computed using their flow and local features.

determines if two or more segments belong to the same group depends on the camera view.

For the first rule, the method select some pixels of the moving segments and uses its corresponding optical flow values in order to determinate the direction of the moving segment. For that propose, we build a histogram of directions composed by eight bins. Each histogram bin counts the frequency of occurrence of pixels pointing in a specific direction. Finally, the bin with the highest value is took as the main direction of the segment.

The second rule is very intuitive: moving segments pointing towards a same direction that are close enough give us the final clue of a possible group of people. To determine whether two segments are close, we measure the distance between the segments, *i.e.*, we want to calculate the distance between the closer sides of the segments. For doing that, we represent the segments using a circular shape, where its centers coincide with the center of the circle. The radius of the circle is equal to half of the longest size of the bounding box. The distance that measures the closeness between two segments is given by the distance of the centers of the circles minus the sum of the radii of the circles.

**Grouping: Local Features.** The first clue for grouping segments is the direction, the segments must be pointing to the same direction, but this occurs in an ideal scenario where occlusions between people are not considered. However, occlusion is a common situation in groups. To overcome this problem, we propose to extract features from segments and let a classifier determine if it contains a single person or a group of people.

To describe each moving segment, we use a variation of the idea proposed in [18], where each segment has temporal information that helps for the comprehension of the movement. The moving segments of each frame are described using HoG [19] descriptor with temporal information. The vector that characterizes each segment is formed by the concatenation of the HoGs extracted from the segments of three consecutive frames. Since segments may vary in size, each one is resized to a fixed size before extracting HoG. Then, the features extracted from each segment are presented to a SVM, responsible for classifying it in containing either a group of people or a single person.

## 4 Experiments and Discussion

In this section, we present the experimental evaluation performed using our proposed approach in two datasets: PETS 2009 and BEHAVE. In addition, we explain the parameters used in it and we discuss the results, especially the success and the mistakes in the detection process. The code was written in C++ using the OpenCV image processing library [20].

The PETS 2009 dataset contains images with $768 \times 576$ pixels and the number of frames varies between 250 to 800, acquired at a frame rate of 14 frames per second. It has four video sequences, the first, $S_0$, contains only background images with almost no people, which is therefore used to train the approach. The remaining sequences $(S_1, S_2, S_3)$ are more complex in terms of number of persons, containing a large number of groups. For the PETS 2009 dataset, we use the view 1. The BEHAVE dataset contains various scenarios of people acting out various interactions. The data was captured at 25 frames per second in a $640 \times 480$ resolution. The authors provide the ground truth, this one was annotated with Viper [21] tool, the same we used to build the ground truth for the PETS 2009. The ground truth of this dataset contemplates situations such as InGroup (IG), Approach (A), WalkTogether (WT), Split (S), Ignore (I), Following (FO), Chase (C), Fight (FI), RunTogether (RT), and Meet (M). However, we only consider IG and WT since our goal is only to detect the existence of groups.

**Results in the PETS 2009 Dataset.** Our model is divided in three steps. The first regards moving object segmentation, the background image is generated using MoG technique, the equalization values are defined by default in OpenCV, the number of frames between each background generation is 14. Since we have the background image, the moving objects are segmented by frame difference technique, close morphological operation is used to smooth the foreground image, the structure element has 7x7. Tiny segments are discarded using a simple threshold that take in to count the number of border pixels of the segment.

Selected segments are tracked using the optical flow technique. For each new frame, segments are paired with the prediction from the previous frame. The segments has a life time, this was used to control if the segment still remains in the frame or if it joints to a group. Thus, too small segments and segments that are leaving out of the frame are quickly eliminated.

The first grouping criterion is easily recognized and is an emulation of how human would bring together a group of people in a single look, using just proximity and same direction. Proximity is measured by the space between the circles generated by the rectangles that contain the segments. The value used in the dataset PETS was set 50 pixels due to the frame size and the distance of humans from the camera. The direction of the segments is calculated using histograms of flow orientations. The general orientation prevails in the histogram except for the moments when the segment is abruptly stopped or changed direction, in such case is not possible to make a correct prediction.

Regarding the second criterion, the SVM classifier was trained with the feature vectors extracted from small sequences of the database segments. The dimension of each feature vector is 11340, resulting from the concatenation of the

current frame with the two previous frames (3780 variables each). The SVM training was performed using 1000 feature vectors.

The PETS 2009 dataset provides no ground truth for groups, then it was necessary to build a ground truth of groups for the image sequences, which was made using the Viper tool [21]. To make the comparison between the ground truth and the output of the model is necessary to make an analysis of the groups found by the model and matching them with the ground truth segments, this process must be done frame by frame.

For pairing groups, an overlapping analysis is performed. The largest bounding box corresponds to segment 1, the other corresponds to segment 2 (*i.e.*, $Seg_1$ could be a group in the ground truth and $Seg_2$ could belong to result of our method). $Seg_1$ is a bounding box represented by $A$ and $C$ points, in same way, $E$ and $G$ represent the segment $Seg_2$. Each point has its $(x, y)$ coordinates. To determinate the superposition area, we find the $I$ and $F$ points. Both can be find with next formules: $I_x = min(A_x, E_x)$, $I_y = max(A_y, E_y)$, $F_x = max(C_x, G_x)$ and $F_y = min(C_y, G_y)$. To finish the matching process, we check if any of the vertices of the area of superposition are within any other segments.

The ground truth generated by us attempted to introduce a more high level definition of group of people. For instance, if two people are sporadically close, they will not be considered as a group, however, if they are a bit distant but keeping a conversation, they will be grouped. It should be noted that even with grouping criteria defined, the decision of being a group was subjective at some times due to the lack of context given the short tracklets available in the video.

The results achieves in the PETS 2009 dataset ($S_2$ and $S_3$) are summarized in Table 1. According to the results, we can see that the first criterion that uses proximity and flow is not as very accurate since occlusion brings people together but they are not in a group. The second one has more success because many of the groups are in a single segment. $S_2$ contains a smaller number of people but confused and walking in different directions, so that a large number of occlusions took place by intersecting people, from there the error in the criteria adopted. In the $S_3$ there are more people than $S_2$. In this case the first grouping criteria achieved a higher rate of success. Finally, the combination of these criteria increases the hit rate in the group detection. This is because a larger number of groups are formed when the we use these criteria. This obviously helps in group detection, but also increases the number of false positives as can be seen in Table 1. This issue will be discuss in the next paragraphs using short sequences.

Table 1: Accuracy for PETS 2009 (left) and false positives for PETS 2009 (right). Results are in %)..

| | Criteria | | |
|---|---|---|---|
| Pets | Flow | Descriptor | Combination |
| $S_2$ | 28 | 75 | 86 |
| $S_3$ | 85 | 90 | 96 |

| | Criteria | | |
|---|---|---|---|
| Pets | Flow | Descriptor | Combination |
| $S_2$ | 64 | 38 | 40 |
| $S_3$ | 5 | 10 | 5 |

In $S_2$, in which can be seen in groups of people with low density, each are separated considerably. In this case, the false positive rate was high because people were constantly in common points, but in different directions. In the frames we can see an intersection of tracks. Despite the considerable rate of false positives, true groups were also detected. In sequence $S_3$ from PETS, the success rate was high due to the conglomeration of people. At the same time the rate of false positives dropped, since the groups were more dense and generally moved in the same direction.

**Results in the BEHAVE Dataset.** For the BEHAVE dataset, we only perform tests on the classifier (second criterion) since the data is not suitable for executing background subtraction or optical flow due to the way it was captured. Using the same ground truth database provides, we determine the groups only using the descriptor-based approach. The classifier was trained with images belonging to PETS 2009 and BEHAVE. The overall success rate for in group and walking together set was 90% in clip 1 – considerably high for these cases.

## 5 Conclusions

In this paper, we proposed an approach to detect groups of people in video sequences. Our method is based in the following steps: moving object segmentation, tracking segment, flow estimation and local feature description. The principal advantage of our method is its simplicity in grouping people. We presented two criteria for group detection. The first is based on proximity and direction flow of persons. These two clues were used for tracking the moving objects and define if they belong to same group. The second is based on temporal HoG features.

According to the experimental results performed in two well-known datasets, the PETS 2009 and the BEHAVE, the combination of both aforementioned criteria outperformed the results provided by them being executed separately. While the first determinate whether two or more segments belong to a group, the second confirms if there is a group in a moving segment even in cases where partial occlusions may take place.

## Acknowledgments

The authors would like to thank the Brazilian National Research Council – CNPq (Grant #487529/2013-8) and the Minas Gerais Research Foundation - FAPEMIG (Grant APQ-01806-13).

## References

1. J. Aggarwal and M. Ryoo, "Human Activity Analysis: A Review," *ACM Comput. Surv.*, vol. 43, no. 3, pp. 1–43, 2011.
2. R. B. F. S. J. Blunsden, "The behave video dataset: ground truthed video for multi-person behavior classification," in *Annals of the BMVA*, vol. 4, 2010.

3. S. Saxena, F. Brémond, M. Thonnat, and R. Ma, "Crowd behavior recognition for video surveillance," in *Proceedings of the 10th International Conference on Advanced Concepts for Intelligent Vision Systems*, ACIVS '08, (Berlin, Heidelberg), pp. 970–981, Springer-Verlag, 2008.
4. B. Zhan, D. N. Monekosso, P. Remagnino, S. A. Velastin, and L.-Q. Xu, "Crowd analysis: a survey," *Mach. Vision Appl.*, vol. 19, pp. 345–357, Sept. 2008.
5. S. Vishwakarma and A. Agrawal, "A survey on activity recognition and behavior understanding in video surveillance," *The Visual Computer*, vol. 29, no. 10, pp. 983–1009, 2013.
6. S. Hongeng and R. Nevatia, "Multi-agent event recognition.," in *ICCV*, pp. 84–93, 2001.
7. M. S. Ryoo and J. K. Aggarwal, "Stochastic representation and recognition of high-level group activities," *IJCV*, vol. 93, no. 2, pp. 183–200, 2011.
8. S. Y. Bingbing Ni and A. Kassim, "Recognizing human group activities with localized causalities," pp. 1470–1477, 2009.
9. Y. Yin, G. Yang, J. Xu, and H. Man, "Small group human activity recognition," in *ICIP*, pp. 2709–2712, 2012.
10. E. Andrade, R. Fisher, and S. Blunsden, "Detection of emergency events in crowded scenes," in *Crime and Security, 2006. The Institution of Engineering and Technology Conference on*, pp. 528–533, June 2006.
11. W. Lin, M.-T. Sun, R. Poovendran, and Z. Zhang, "Group event detection for video surveillance," in *Circuits and Systems, 2009. ISCAS 2009. IEEE International Symposium on*, pp. 2830–2833, May 2009.
12. S. Saxena, F. Brémond, M. Thonnat, and R. Ma, "Crowd behavior recognition for video surveillance," in *Proceedings of the 10th International Conference on Advanced Concepts for Intelligent Vision Systems*, ACIVS '08, (Berlin, Heidelberg), pp. 970–981, Springer-Verlag, 2008.
13. S. Kwak, B. Han, and J. H. Han, "Multi-agent event detection: Localization and role assignment," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pp. 2682–2689, June 2013.
14. M.-C. Chang, N. Krahnstoever, S. Lim, and T. Yu, "Group level activity recognition in crowded environments across multiple cameras," in *AVSS*, pp. 56–63, 2010.
15. I. Rodrigues de Almeida and C. Rosito Jung, "Change detection in human crowds," in *Graphics, Patterns and Images (SIBGRAPI), 2013 26th SIBGRAPI - Conference on*, pp. 63–69, Aug 2013.
16. T. Bouwmans, F. E. Baf, and B. Vachon, "Background modeling using mixture of gaussians for foreground detection a survey," in *Recent Patents on Computer Science*, pp. 219–237, 2008.
17. J. yves Bouguet, "Pyramidal implementation of the lucas kanade feature tracker," *Intel Corporation, Microprocessor Research Labs*, 2000.
18. Y. Zhang, X. Liu, M.-C. Chang, W. Ge, and T. Chen, "Spatio-temporal phrases for activity recognition," in *ECCV*, ECCV'12, pp. 707–721, 2012.
19. N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1, pp. 886–893 vol. 1, June 2005.
20. G. Bradski, "Opencv library," *Dr. Dobb's Journal of Software Tools*, 2000.
21. D. Doermann and D. Mihalcik, "Tools and techniques for video performance evaluation," in *Pattern Recognition, 2000. Proceedings. 15th International Conference on*, vol. 4, pp. 167–170 vol.4, 2000.