A Smart Camera Network with SVM Classifiers for Crowd Event Recognition

Jhih-Yuan Huang Wei-Po Lee

Abstract—**It is now popular to develop video surveillance systems to enhance the security of our daily life. At present, the surveillance systems have turned to automatically identify the continuous human behaviors to detect various events from video streams. It means that more efficient approaches are in demand to deal with the largely increasing amount of data. This study presents a low-cost distributed smart cameras system, together with a machine learning technique to detect abnormal events by analyzing the sequential behaviors of a crowd of people. Moreover, this system employs a collaboration strategy to perform collective decision making for event recognition. Experiments have been conducted to evaluate the proposed approach and the results confirm its reliability and stability in event recognition.**

Index Terms—**video surveillance system, smart camera network, collective intelligence, support vector machine, event recognition**

I. Introduction

It is now popular to develop video surveillance systems to enhance the security of our daily life. With the aid of video streams recorded by the surveillance equipment, security staffs can detect sudden unusual events and respond promptly to the emergent situations rapidly to reduce the risks. To detect various events from video streams, the surveillance systems have now turned from the analysis of individual images to the continuous human behaviors [1][2]. In many cases, a single view is not sufficient enough to cover a target region network, and a network of cameras is thus required to cope with an open area in which many people move arbitrarily. At present, most of the camera networks follow a centralized architecture, which often suffers the problems of high communication cost and scalability [3]. Therefore, an efficient surveillance system needs not only to perform behavior recognition, but also to overcome the problem of bandwidth limitation. A promising solution is to adopt a distributed smart camera sensor network [4]. The camera nodes can process locally available images, perform data compression, and transmit the results to the neighbor nodes in the same network for information sharing. The nodes communicate in a peer-to-peer-manner and only abstract information is exchanged between nodes. In this way, the overall computation can be achieved in a distributed way by a set of inexpensive devices.

The main goal of developing a distributed smart camera system in public area is to recognize human behaviors for abnormal event detection. The abnormal events here mean the observable events that occur unexpectedly, abruptly and unintentionally, and they invoke an emergency situation that requires fast responses [5]. To achieve the above goal, some important issues need to be considered. One is to extract pedestrian features from the video streams recorded by the cameras. With a set of properly defined features to represent target data, a computational method can thus be employed to construct robust and reliable classifiers for event recognition. The other issue is to build behavior sequences with the selected feature. Though there have been many works focusing on how to precisely construct behavior sequences for the pedestrians from different image frames, most of such approaches are expensive in computation. To deploy the distributed camera network approach to a real life environment, more simple and efficient strategies are needed. The data representation for these persons must be concise to ensure the efficiency and effectiveness of the learning method. Thus, the traditional encoding scheme of combining all personal features from the crowd is not preferred and a new encoding scheme is needed. In addition, individual cameras are often not able to capture complete behavior sequences perfectly, due to some environmental factors in the real world, such as the blind angles of the camera network. An efficient strategy with a relatively low resource need (in terms of computing and communicating) is required to exploit the device collaboration within a smart camera network.

To overcome the difficulties associated with the above issues, in this work we develop a distributed smart camera network system with several unique features. They mainly includes a simple but efficient strategy to organize the behavior sequence, a new indirect encoding scheme to represent a crowd with relatively few features, a machine learning approach to train robust event classifiers, and a collaborative strategy to detect abnormal events by collective decision. To evaluate the performance of the proposed approach and to compare different strategies coupled within the camera network, a series of experiments have been conducted. The results confirm the efficiency and effectiveness of the proposed approach.

II. Related Work

The video sensing techniques have been widely applied to different surveillance systems. Different approaches for video-based activity recognition has proposed and implemented. These approaches differ mainly in the underlying image sensing and processing techniques (such as

J.-Y Huang and W.-P. Lee are with Department of Information Management, National Sun Yat-sn University, Kaohsiung 80424, Taiwan (e-mail: wplee@ mail.nsysu.edu.tw).

motion detection and feature extraction), and the machine learning methods (such as naive Bayes classifiers, hidden Markov models) adopted to build the recognition models [2][6].

In the video-based recognition studies, Govindaraju *et al.* defined a three levels semantic hierarchy to describe what occur in a video [7], including action, activity and event. Considering the hierarchical relationship between action, activity, and event, the main focus of the surveillance system now moves towards the event detection and recognition, after the goals of recognizing simple body movements and human activities are achieved. A video event can only be understood through a sequence of images. Therefore, analyzing video data for event recognition is an inherently time consuming task, due to the streaming nature of the data. Some approaches have been proposed to extract representative features for image frames to reduce the data amount in the recognition procedure. For example, in [8] the authors combined both the differential and template approaches to compute the frame-to-frame difference and extract shape moment descriptors with the temporal motion trajectory. To obtain even more accurate recognition results, some researchers have exploited the advantages of multiple cameras to infer human activities: the authors in [9] used multiple cameras to build a system for human 3D activity recognition. More detailed works on event recognition are referred to [10].

Recognizing abnormal events from normal ones is especially important in the study of event recognition. It can be applied to various real life surveillance applications [11] [12]. For example, in [13], the authors developed an event detection framework for elderly healthcare, and the smart cameras have been used to detect a human falling in a home environment. However, previous studies tended to infer events from the single person's behavior, but hardly extracted and analyzed behavior data of a crowd of people. The analysis of several people needs more advanced cameras and introduces higher error rates, due to the behavior discrepancy of the people being observed. Our approach involves a set of low cost smart cameras to make the recognition decision in a collective way. Further details are described in the sections below.

III. A Distributed Smart Camera Network for Collective Event Recognition

Our main goal is use a low cost camera network for abnormal event detection. To achieve this goal, we develop an approach that takes the advantages of collective intelligence: a high-level phenomenon that emerges naturally from the interplay of collaboration and competition of many individuals of a population [14][15]. It is usually defined as the ability of a group to solve problems which cannot be solved by the members individually.

Figure 1 illustrates the system architecture, the core units of the system, and how the system operates. The system flow includes the following steps: (1) the raw image data are captured from the video sensor; (2) the regions of interest are marked and the pedestrians are detected; (3) the pre-defined features are extracted from the marked regions and the successive features for each pedestrian are recorded; (4) features are reduced and encoded to form a data record, and feature data from different persons in the same region are combined to be a data vector (for training); (5) classifiers are built from the training data and used to detect patterns of abnormal events; and (6) each camera goes through the communication unit to transmit its processed data to the neighbor cameras, and pass the result of event recognition to the management center. Each camera performs the steps described above and they work together to achieve the overall task of event detection in a surveillance area. The details of the core units are described in the following subsections.

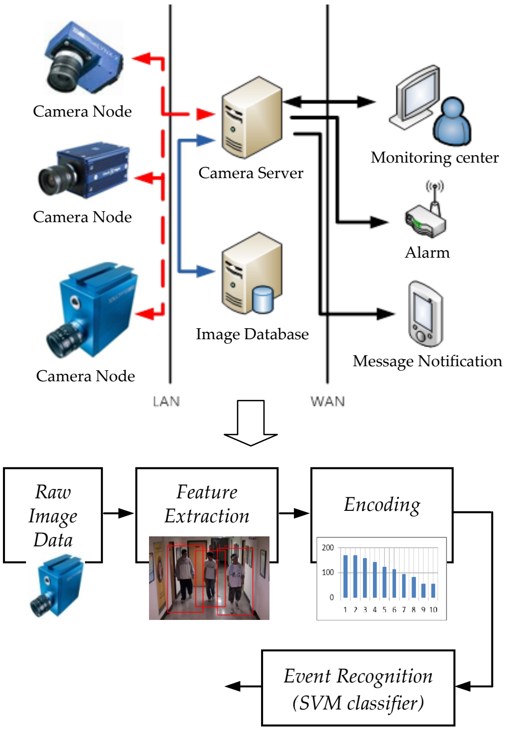

Fig. 1. The core units of our collective recognition system.

A. *Feature Extraction and Behavior Sequence Collection*

The first phase in event detection is to extract specific target features from the video stream and then the system can infer what event is happening accordingly. To reduce the computational load of the cameras, our approach simply samples some frames within a specific time interval. In this work, we adopt the OpenCV (http://opencv.org/) to analyze the Histogram of Oriented Gradients (HOG) to determine if there is any person existing in an image.

With the cost constraint (note that the goal is to develop a low-cost smart camera network for event recognition), it is essential to select a proper feature to achieve a reasonable system performance in real time. Here, the system draws a region of interest (ROI, the rectangle in the image of Figure 2, meaning the current height of a person in the image) for each person detected. Then, we take the height of ROI, as the feature data to represent a person and develop a mechanism to use the changes of ROI recorded from an event duration to constitute a feature vector for event recognition.

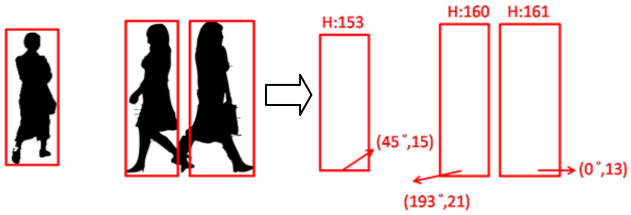

Fig. 2. Extracting ROI from the original image.

To detect a specific event, the system needs to continuously detect and mark the pedestrians, and then extract the feature data described above from their behavior sequences. Because several people may move within the surveillance area a smart camera can cover, the system has to make distinctions among them to construct a correct data sequence for each person. One plausible way is to compare the HOG of two consecutive images to ensure the consistency of the targets in the two images (i.e., belonging to the same behavior sequence). However, this is more computationally expensive than a single smart camera can afford. Therefore, we design a simple but efficient strategy that can reduce the number of the comparisons to save the computational cost. Here, we assume that human behaviors are harmonious and inertial, so the directions of a person's body parts are often consistent (or not changed suddenly in very short period of time) in motion. With this assumption, the system only performs the HOG comparison for the images to build a forward vector for a person when he first appears in the space a smart camera is monitoring. After that, it calculates the cosine value of two consecutive motion vectors but not the complete HOG computation to confirm the moving trajectory of a single person (by taking the image with minimal cosine value).

In addition to identifying the behavior sequence for each person in the same surveillance area, the system has to convert the height of each ROI to its original size (the discrepancy is caused by the distance between the target and the camera in each image: a larger size ROI is obtained when a person moves towards the camera), in order to construct a complete feature vector for each person. This problem can be solved by using the camera model theory that exploits the principle of perspective projection of pinhole imaging. Assume that the focal length of a camera is known, the well-known camera model can be used to calculate the position and the size of an object projected in an image. On the contrary, this process can be reversed: if the position and size of an object are known, the original width and height of an object can be estimated by the same model. Therefore, we directly use the above method to convert the feature values to its original form for the series of image frames collected from the procedure of behavior sequence identification described above.

B. *Feature Data Encoding for a Crowd of People*

As mentioned above, once a person's behavior sequence is identified and the ROI sizes are converted, a feature vector for a single pedestrian can be built. To detect a specific event in a public space, the system needs to observe the behaviors of a crown of people in the same region (rather than that of any individual), and infer what event is happening accordingly. A direct way is to append the feature vectors built for different persons to form a new vector for the crown. But it should be noted that the dimension of this combined vector will increases dramatically along with the number of people in the surveillance area. The high dimensional feature data thus become a burden for a smart camera network: though with limited computational power, it is expected to response to an abnormal event in real time. Under such circumstance, the dimension of the combined feature vector has to be further reduced.

In this work, we develop an indirect encoding scheme that derives a concise and compact representation from the feature vectors described above to represent the behavior sequences of a set of persons. In this representation, the vector for each person is reduced to include three new features: the maximal change of the ROI height, the number of image frames within which the maximal change happens, and the frequency of the considerable ROI change. The first feature is to measure the maximal behavior variation (i.e., the difference of the maximal and minimal ROI heights; the value could be positive or negative) of a person within a pre-defined time interval (i.e., ten time steps in our experiment). This value is normalized, subject to the maximal height. The second feature means to provide the changing rate from the maximal/minimal to minimal/maximal heights in terms of the number of image frames. And the third feature reports how often a person changes his behaviors by measuring the variation (i.e., the slope) between two consecutive ROI heights and then checking if the variation exceeds a pre-defined threshold. In this way, the dimension of the combined feature vector can be largely reduced and the system performance can thus be improved.

C. *Collective Event Recognition*

After encoding the behavior sequences of a crown of people as feature data, we adopt a machine learning method to classify the target events occurring in a surveillance area. According to several related studies (e.g., [16]), support vector machine (SVM) is the method most suitable for our classification task here, due to its good performance in dealing with multi-class and high-dimensional data constituted by real numbers. Therefore, we choose to use SVM classifiers for event detection in this work.

To train a SVM classifier, the system includes the offline training and online operating phases. The training phase involves collecting historical video files and analyzing of video streams for features data extraction. The data can be processed on a workstation or a server level computer (rather than on a smart camera) of which its computation is more powerful to build a classifier. The classifier built can then dispatched to all smart cameras in the same network. In this work, we use the online available software LiSVM ([17]) with a kernel type of linear to construct the SVM classifiers. This configuration is selected as it achieves a good balance of the processing speed and the recognition accuracy.

The recognition phase operates on each single smart camera. As mentioned, a smart camera has an embedded system structure with very limited computational power and storage (compared to the traditional personal computers or central servers). Therefore it only functions for feature extraction, encoding, and event recognition. In addition, each camera can duplicate the behavior sequences it has identified and transmit them to the neighborhood sensor nodes to ensure the completeness of each feature vector (i.e., to mend any incomplete data). Though the high level computer vision algorithms can be divided so that intermediate results can be exchanged with other cameras, we do not perform such partition because our goal is to use relative small amount of shared information to achieve event recognition through a collective decision making strategy. Therefore, in our approach, a smart camera just transmits the processed information unless an abnormal event is detected. In this way, the transmission load of the system can be largely reduced and the system can thus be scaled up to include more sensor nodes to monitor a huge surveillance area.

IV. Experiments and Results

A series of experiments has been conducted to evaluate the proposed approach for event detection. The target events include earthquake, gun shooting, and fighting. For practical reason, a simulation-based strategy was adopted for data collection. Three scenarios were designed for the above events and some participants were asked to demonstrate their responses to different events. The collected data samples were then extended to generate data sets through a statistic-based approach.

The creation of simulated data for the earthquake event was based on the observation that the possible crowd behaviors have a common feature: the height of a person was decreasing, though the moving speed and distance of different persons might differ from one to another. Thus, to create a data set to train classifiers for this event, we collected the sample behaviors from the demonstration procedure and used these samples to build a normal distribution to randomly generate a height variation for each person at each time step (as shown in Figure 3). That is, in simulation the height of a person varied according to the rule $H_{t+1} = H_t - N(\mu,\sigma)$, where H represented the height and N was a normal distribution with mean μ and standard deviation σ. Similarly, two datasets were generated for the events of gun shooting and fighting respectively. The rule used for the gun shooting was the same as the one for earthquake, while the rule for the fighting event was changed to $H_{t+1} = H_t + N(0,\sigma)$. This distribution has a mean 0 in order to produce positive and negative height variations for a person when he was fighting with others.

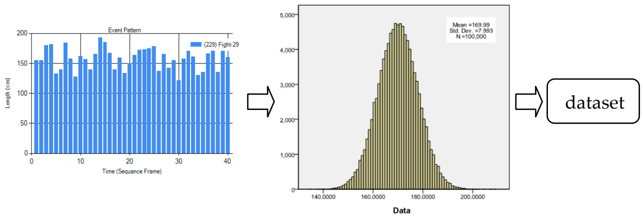

Fig. 3. Creating datasets for the experiments.

A. *Single Event Recognition*

With the above data sets, different sets of experiments were conducted to verify the feasibility and reliability of our approach. As described in section 3.1, each data record is constituted by the features collected from several persons (ten in this work). It is notable that people may have different responses to the same event. To investigate the effect of inconsistent human behaviors, in the first set of experiments, we designed several strategies with different ratios (ranging from 0.0 to 1.0) of two types of behaviors (i.e., target or non-target) to train and test the SVM classifiers. For example, in the case of earthquake, a ratio of 0.1 means that there were 10% of people in the crowd with non-target behaviors (which were randomly selected from distributions of gun shooting and fighting), and 90% of people with target behavior (selected from the distribution of earthquake). For each strategy (i.e., ratio), 1000 data records (each record included ten people) were created.

The SVM classifiers were built for single event recognition (i.e., to predict a data record is a target event or not), in which a 10-fold cross validation method was used to train and test the classifiers. The results (i.e., accuracy) for each strategy are presented in Figure 4. In addition to the training phase, different ratios were also used in the test phase to observe the corresponding effects. As can be seen in this figure, the classifiers trained from feature data with certain levels (i.e., 0% up to 60%) of non-relevant behaviors, are robust and able to perform precise recognition in the test phase. But the performance declines (i.e., the right hand side of the figure) when the test data contained more than 50% of non-target behaviors in each crowd (i.e., ten persons) data. On the contrary, the classifier built from data with a high ratio of non-target behaviors (i.e. 80%) could not deliver a high recognition accuracy as others for the test cases with low ratios of non-target behaviors, while better results were obtained for the test cases with high impurity (in which the test data became more and more similar to the training data). This is mainly due to the situation of class imbalance. Therefore, to obtain a higher accuracy, the amount of different classes of training data needs to be arranged carefully.

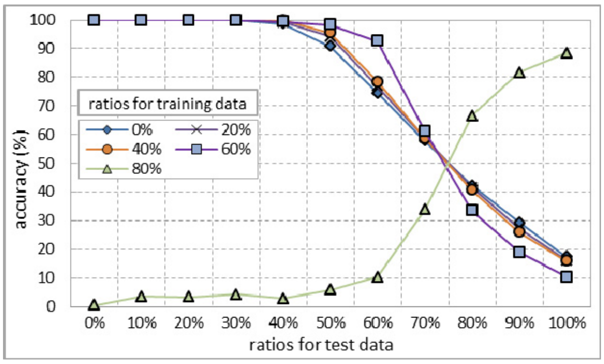

Fig. 4. The experimental results for strategies with different levels of data impurity.

B. *Multiple Events Recognition*

After the series of experiments of building SVM classifiers for single event detection, we employed the same approach to train classifiers for multiple events recognition. In this experiment, the output of the classifier needed to indicate which event among the four (earthquake, gun shooting, fighting, and normal) was happening. The same training and test procedures as above were performed and the results are shown in Figure 5. As can be observed, the effects caused by different ratios are similar to those obtained from the single even recognition presented in Figure 4, though the classifiers for multiple even recognition are not as precise as those for single event recognition. It is worth noting that classifiers trained from the data with lower levels impurity (i.e. 0~20%) did not have results with high accuracy in the test cases with high rations of data impurity, because the classifiers have overly fitted the training data.

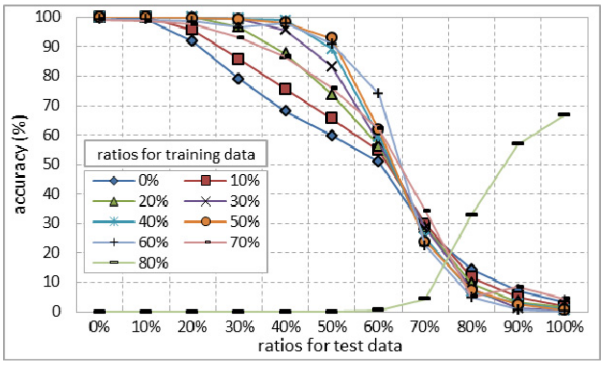

Fig. 5. Results for multi-event recognition.

C. *Collective Event Recognition*

In the real world applications, it is in fact difficult for the collective smart cameras to capture all behavior changes among a crowd of people, due to some undesired environmental effects (such as the blind angles of mounted sensors and coverage rates of the sensors). That is, the image frames collected from the cameras may be incomplete and thus not enough to form a behavior sequence. To overcome such as a problem, the system needs some strategies amended and to achieve precise event recognition in the real life environment. One possible way is to distribute the images collected by a camera to its neighbors to repair the behavior sequence. Though there have been many algorithms proposed to construct complete data, however, they are mostly computationally expensive. Therefore, here we propose to recognize the event through collective decision, in which each camera perform the recognition task independently and all cameras voted for the final decision.

To evaluate the collective decision method, a series of experiments was conducted. In the experiments, the behavior sequence for each single person (including 15 consecutive image frames as described in section 3.1) was randomly impaired up to 20% (i.e., 1~3 frames were removed from the original data) to simulate the effect of incomplete data. In addition, a ratio of 0.3 was used to train SVM classifiers (as it was reported previously to derive a robust classifier). Four different numbers of smart cameras were tested for the collective surveillance, and the results are presented in Figure 6 They show that incomplete data caused different degrees of damage in recognition on different types of event. As can be seen in the figure, when more cameras were deployed in the environment with a collective decision, the performance can be improved effectively.

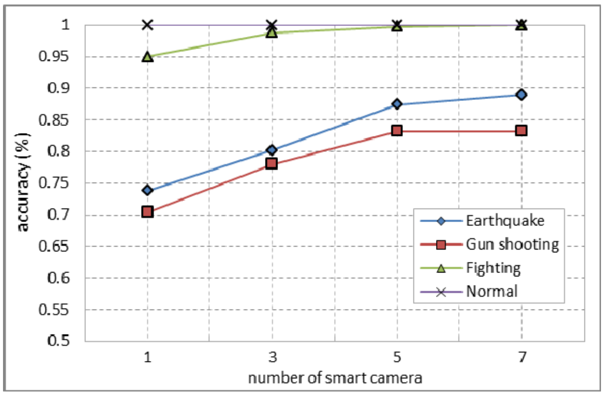

Fig. 6. Results of collective surveillance.

V. Conclusions and Future Work

To deploy a surveillance system in a cost efficient manner, in this work we presented a simple but efficient strategy on a smart camera network to organize human behavior sequence, and a new encoding scheme to represent a crowd. Most importantly, we developed a collective approach for the detection of abnormal events. In our approach, the smart cameras can cooperate with each other for making group decision. Different series of experiments have been conducted to evaluate the developed approach, and the results show that our system can provide a certain degree of stability and reliability with a low cost.

Currently, we are investigating new ways to address the scalability issue of the camera network. Also, we plan to develop an adaptive mechanism to detect the technical faults of the cameras, and automatically adjust the visual angles of some neighborhood cameras to monitor the area where faulty cameras could not execute their functions.

References

[1] K. Popoola and O. P. Wang, "Video-based abnormal human behavior recognition-a review," *IEEE Trans. Systems, Man, and Cybernetics*, Part C, vol. 42, no. 6, pp. 865–878, 2012.

[2] N. C. Krishnan and D. J. Cook, "Activity recognition on streaming sensor data," *Pervasive and Mobile Computing*, vol. 10, pp. 138-154, 2014.

[3] V. P. Munishwar and N. B. Abu-Ghazaleh, "Scalable target coverage in smart camera networks," in *Proceedings of the Fourth ACM/IEEE International Conference on Distributed Smart Cameras*, pp. 206-213, 2010.

[4] B. Song, A. T. Kamal, C. Soto, C. Ding, J. A. Farrell, and A. K. Roy-Chowdhury, "Tracking and activity recognition through consensus in distributed camera networks," *IEEE Trans. on Image Processing*, vol. 19, no. 10, pp. 2564–2579, 2010.

[5] M. J. Roshtkhari and M. D. Levine, "Online dominant and anomalous behavior detection in videos," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[6] T.-K. Truong, C.-C. Lin, and S.-H. Chen, "Segmentation of specific speech signals from multi-dialog environment using SVM and wavelet," *Pattern Recognition Letters*, vol. 28, vol. 11, pp. 1307-1313, 2007.

[7] G. Venu, "A Generative framework to investigate the underlying patterns in human activities," in *Proceedings of IEEE International Conference on Computer Vision Workshops*, pp. 1472–1479, 2011.

[8] W.-C. Cheng, “PSO algorithm particle filters for improving the performance of lane detection and tracking systems in difficult roads,” *Sensors*, vol. 12, pp. 17168–17185, 2012.

[9] M. B. Holte, C. Tran, M. M. Trivedi, and T. B. Moeslund, “Human pose estimation and activity recognition from multi-view videos: comparative explorations of recent developments,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, no. 5, pp. 538-552, 2012.

[10] N. S. Suriani, A. Hussain, and M. A. Zulkifley, “Sudden event recognition: A survey. *Sensors*, vol. 13, pp. 9966-9998, 2013.

[11] D. Lymberopoulos, T. Teixeira, and A. Savvides, “Macroscopic human behavior interpretation using distributed imager and other sensors,” *Proceedings of the IEEE*, vol. 96, no. 10, pp. 1657-1677, 2008.

[12] G. Singla, D. J. Cook, and M. Schmitter-Edgecombe, “Recognizing independent and joint activities among multiple residents in smart environments,” *Journal of Ambient Intelligence and Humanized Computing*, vol. 1, no.1, pp. 57-63, 2010.

[13] A. Bamis, D. Lymberopoulos, T. Teixeira, and A. Savvides, “The BehaviorScope framework for enabling ambient assisted living,” *Personal and Ubiquitous Computing*, vol. 14, no. 6, pp. 473-487, 2010.

[14] A. W. Woolley, C. F. Chabris, A. Pentland, N. Hashmi, and T. W. Malone, “Evidence for a collective intelligence factor in the performance of human groups,” *Science*, vol. 330, no. 6004, pp. 686-688, 2010.

[15] D. Król and H. S. Lopes, “Nature-inspired collective intelligence in theory and practice,” *Information Sciences*, vol. 182, pp. 1-2, 2012.

[16] G. Lavee, E. Rivlin, and M. Rudzsky, “Understanding video events: a survey of methods for automatic interpretation of semantic occurrences in video,” *IEEE Trans. on Systems, Man, and Cybernetics*, Part C: Applications and Reviews, vol. 39, no. 5, pp. 489-504, 2009.

[17] C.-C. Chang and C.-J. Lin, “LIBSVM: a library for support vector machines,” *ACM Trans. on Intelligent Systems and Technology*, vol. 2, no. 3, pp. 1-27, 2011.