This CVPR2014 Workshop paper is the Open Access version, provided by the Computer Vision Foundation.
The authoritative version of this paper is available in IEEE Xplore.

Ground-Based Activity Recognition at Distance and Behind Wall

Tao Wang*, Riad Hammoud*, and Zhigang Zhu+

* BAE Systems, 6 New England Executive Park, Burlington MA 01803
{tao.wang | riad.hammoud}@baesystems.com
+ Department of Computer Science, The City College of New York, New York, NY 10031
zhu@cs.ccny.cuny.edu

Abstract

Long-range activity recognition is a challenging research problem in a surveillance area where sensors cannot be placed close to targets. Even a simple activity can be confused with other activities or not be recognized correctly if the detection in one of the sensor modalities is not certain or even unavailable. Also, the training of some real-life activities is not feasible, because it is hard to collect sufficient and accurate labeled data for varieties of free-living activities. In this paper, we use an unsupervised learning algorithm, Dirichlet process Gaussian mixture model (DPGMM), to construct a model to determine the number of classes automatically. To further represent a set of features as one event, and communicate between both audio and video, we use the DPGMM as a base and enhance it with additional aggregation, multimodal association and transition. This new model is called aggregation coupled Dirichlet process Gaussian mixture model (AC-DPGMM). We present experiments with some activities that cannot be simply distinguished using visual features only. Along with audio information, we can also recognize some activities invisible in video, such as speaking behind a wall. We compared our model with a generative clustering algorithm and the original DPGMM, and showed that we have 23.6% and 18.8% improvement in accuracy compared with manually labeled data.

1. Introduction

Activity recognition using both audio and video has been drawing growing interest in surveillance applications [1] [2]. In most activity detection systems [3] [4], only visual information is used. However, in some situations, for example, a person speaking or doing some activities behind a wall that cannot be detected from a visual sensor, audio may be used to convey more information. In the past, microphones have been employed in audio-visual surveillance, but they have the limitation of very short ranges. Also, these sensors need to be placed near a monitoring area. This is particularly a problem when sensors cannot be placed close to a target, especially in a restricted area, or if one sensor modality is unavailable at the time the activity needs to be recognized. A Laser Doppler Vibrometer (LDV), on the other hand, is a long-range, non-contact acoustic measurement device to detect the speed of the target's vibration based on Doppler frequency shift that can be used to obtain the acoustic signals of a sounding target at a large distance. This sensor is well-studied for long-range remote hearing [5] for the purpose of moving vehicle detection [6]. In this paper, we use this type of sensor for long-range audio detection along with a video camera for activity detection at a large distance.

Activity recognition aims to recognize normal and special events. Normal events, such as speaking or walking, need to have a significant amount of training data for accurate parameter estimation. However, this requirement often cannot be satisfied. Special events, such as footsteps or screaming, rarely occur in a consistent timely manner. Hidden Markov Model (HMM) and Gaussian Mixture Model (GMM) are commonly used to recognize activities using an unsupervised machine learning approach [7]. However, they need to have an appropriate number of states pre-defined. In practice, this number is difficult to be known beforehand. A Dirichlet process mixture model (DPMM), on the other hand, can have an infinite number of states and decide the optimal number of states automatically [8]. The DPMM will be the base of our work in this paper. In the learning process, we first extract audio samples into spectral and perceptual features, and video samples into statistical visual features. Since each event may consist of many features that might be clustered into different classes, we need to aggregate them to represent one activity. Also, events from both audio and video need to communicate and be fused properly, especially if information about some activities is not distinctive using one of the sensor sources. For example, in Fig. 1, a person moving a chair (middle) and carrying a heavy object (right) look very similar at a long distance. There is no clear separation using only visual features. Along with audio features, we can separate them as two different activities. For these reasons, we propose an aggregation coupled Dirichlet process Gaussian mixture model (AC-DPGMM) that could improve activity recognition using both audio and video at a large distance. In the experiments reported in this paper, we mainly focus

Approved for Public Release; Distribution Unlimited. Cleared for Open Publication on 4/03/14.

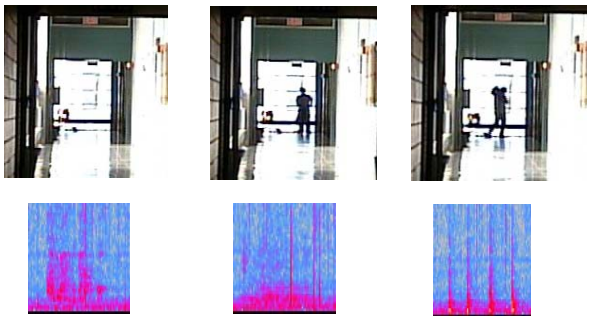

Figure 1: Left column shows a cropped video frame and a truncated audio clip in spectrogram of a person speaking behind a wall. Middle column shows a person moving a chair (middle). Right column shows a person carrying a heavy object (right)

on three activities: a person speaking behind a wall, a person moving a chair in and out of a scene, and a person carrying an object stepping in and out of the same scene. We will compare our AC-DPGMM with a generative clustering algorithm, such as k-means, and the original DPGMM. Performance is evaluated on the feature space with manually truth labeling to compare the results.

The rest of the paper is organized as follows. Section 2 introduces some related work on activity recognition and gives a brief overview of the original DPGMM. Section 3 describes feature selection for both audio and video. Section 4 describes the aggregation coupled DPGMM in general. Then experiment results are provided in Section 5. Finally, conclusions and future work are given in Section 6 and 7, respectively.

2. Related Work

In a multimodal surveillance system, the goal is to detect and identify interesting and/or suspicious activities from multiple sensor sources. Several supervised or unsupervised classification techniques are employed to make decisions among various events. Authors in [9] use spatio-temporal information to enhance the activity recognition process performance. Another way to detect people in a scene is presented by [10], which combines SVM trained Histogram of oriented gradients (HOGs) feature's extracted from color and infrared images using disparity-based detectors. The training of usual events and unusual events is always a hard task. The study in [11] shows that the distinction between two unusual events can be as large as those between unusual events and usual events. The training of a general model for the unusual events is not often feasible. Also, the collection of sufficient amounts of labelled data for a various set of free-living activities may be hard to accomplish and also computationally expensive.

Unsupervised classification techniques, on the other hand, try to build models from unlabeled data by clustering. Several well-known approaches are k-Means algorithm [12], Gaussian Mixture Models (GMM) [13], and Hidden Markov Model (HMM) [14]. HMM is commonly known as a state-space based model that treats the activities and goals as hidden states, and tries to infer such hidden states using statistical learning. However, some activities, such as footsteps, may occur for a very short time. The difficulty might become more significant during the training phase due to the lack of labeled sequences. The fundamental problem of HMM is that it needs to define an appropriate number of states. DPMMs [15] are cast as infinite mixture models to find the correct number of mixture components using the base distribution by assuming infinitely many components. A very recent paper [16] compares DPMM with other algorithms that are commonly used for mixture models (such as expectation-maximization) and shows several advantages in preventing the over-fitting problem and determining the number of components automatically. Simply speaking, Dirichlet process is defined by two parameters, a positive scalar value α and a probability measure G_0, referred to as the concentration parameter and the base distribution, respectively. It is commonly known as a distribution over distributions. It can measure a random probability of a set of samples that may belong to a certain class. In this work, we use Gaussian Mixture Models as the base distributions G_0.

3. Audio/Visual Features

In order to recognize activities from both audio and video, various types of features are extracted from both sensor sources, so that DPGMMs can be applied in the feature space. For every video frame, we perform foreground detection after learning the background information for a couple of seconds. Then, the HOGs [17] feature is used to detect a person. HOG is a statistical feature that preserves some texture and local structure that counts occurrences of gradient orientation in localized dense grid cells uniformly. Since it uses local contrast normalization, it is invariant to illumination changes, thus it is good at detecting people. Principal component analysis (PCA) is then applied to determine representative components and reduce the feature dimensionality.

To synchronize with a visual feature, an overlapped sliding window is applied to raw audio signals corresponding to a particular video frame. The audio features consist of three spectral features and one perceptual feature. The spectral features represent spectral moments and flatness that would be good to detect special activities, such as moving a chair and footsteps while carrying a heavy object. Those spectral features include:

1) Spectral energy: $E = \sum |F\{x(t)\}|^2$, where x(t) is the audio signals and F{} is the Fourier transform. It is used to calculate the energy of the power spectrum.

Non-Technical Data - Releasable to Foreign Persons.

2) Spectral entropy: $-\sum|F\{x(t)\}|\frac{\log|Fx(t)|}{E}$. It is used to measure the energy changes.
3) Spectral flux: $E=\sum|F\{x(t)\}|-\sum|F\{x(t-1)\}|^2$. It is used to measure how quickly the power spectrum of a signal is changing.

The perceptual features represent the spectral variation and sharpness. Mel-frequency cepstral coefficients (MFCCs) [18] are commonly used to perceptually represent the frequency band responses of the human auditory system, thus they are good at recognizing speech.

4. Aggregation Coupled DPGMM

In our study, the first important thing to do is feature aggregation over time. Because DPGMM is applied in the feature space and each event may contain several features representing different classes, we need to aggregate features in a period of time into one activity. This is especially important for special events with sharp and short changes, for example, a footstep event where a person carrying a heavy object may contain a strong sound of footsteps followed by the sound of a silent background if the person moves slowly.

The second important issue is modality association and/or transition. When we have information from more than one source, such as that from both video and audio, they need to be associated to represent similar activities. However, sometimes some activities cannot be detected by one of the sensor sources; for example, there is no visual clue for an event of talking behind a wall. Therefore we need to transit information from one available sensor modality to the next available sensor modality.

With these two consideration, we build our aggregation coupled DPGMM. Figure 2 shows the processing flow to recognize activities based on aggregation coupled DPGMM using audio and visual features. The immediate blocks below audio and above video are original DPGMM, which clusters features into some undefined classes. Let $OE_k=\{Of(t_1)..Of(t_n)\}$ be an event of a class k based on a feature set f in a time period t_1 to t_n. O is a sensor source that can be A for audio, or V for video. The aggregation step is to try to determine a class k that gives the best decision probability of an event, and is defined as:

$$Agg(OE)=\max_k\sum_i^F\alpha\frac{p(f_i|k)p(f_i|\theta)}{p(k,\theta|f_i)},k\in K'\subseteq K$$

where F is the number of features in time period t_1 to t_n. K' is a new estimated number of classes. K is the number of classes that is unknown. It is determined from the original DPGMM which has a positive scalar α and hyper parameters θ that can be estimated via iterations. Note that the new estimated number of classes K' after aggregation will be smaller than or equals to the K from DPGMM. After aggregation, the fusion between audio and video results is done at two separate steps: association and transition. The association is performed when both audio and visual information is available. In this step, the decision probabilities from both results are combined first, then the class k is selected based on the best overall probability. The transition is performed when one sensor's information is unavailable. In this step, the decision probability of the current event is combined with decision probability at the next event using non-background features extracted from a different sensor. Then the best overall probability from two consecutive events is selected using the class k. Note that if such information from another sensor is still unavailable, the decision is simply made based on the current aggregation results.

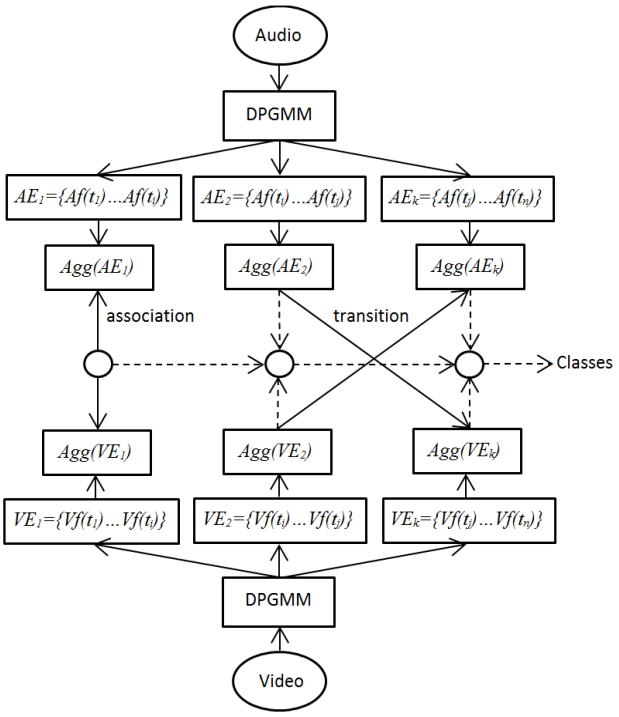

Figure 2: Processing flow chart of AC-DPGMM for a sequence of features

5. Experiment Results

This experiment contains samples collected from a long range Laser Doppler Vibrometer (LDV) and a camera placed at about 420 feet away from a monitoring area, which is the end of a long corridor with a hidden area behind a wall. The laser beam of the LDV is pointed to a metal box on the side wall, which vibrates with acoustic energy created by any close-by activities that produce sound. The events we would like to identify in our experiments include: 1) a person speaking behind the wall, 2) a person moving a chair in and out of a monitoring area, and 3) a person carrying a heavy object stepping in and out of the same monitoring area. The video data is collected at 30Hz, where the audio is collected at

Non-Technical Data - Releasable to Foreign Persons.

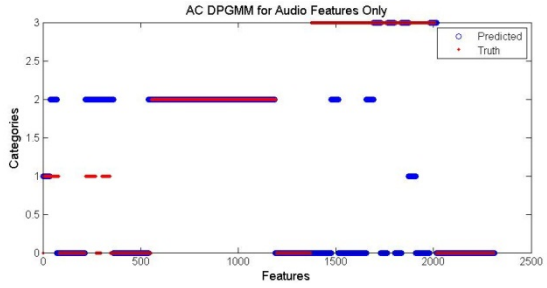

Figure 3: AC-DPGMM for audio features only

Table 1. Confusion matrix of 4 activities using audio only

Acc = 69.01%	B-background, S-speech, C-chair, F-footsteps			
	B	S	C	F
B	789	1	59	6
S	8	35	134	0
C	6	0	635	0
F	392	36	72	138

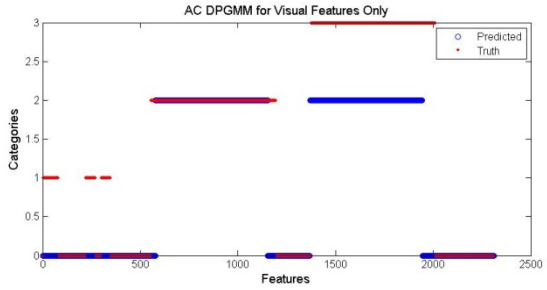

Figure 4: AC-DPGMM for visual features only

Table 2. Confusion matrix of 4 activities using video only

Acc = 61.75%	B-background, S-speech, C-chair, F-footsteps			
	B	S	C	F
B	851	0	4	0
S	177	0	0	0
C	65	0	576	0
F	66	0	572	0

22.5 KHz. A HOG based statistical feature of dimension 243 is extracted for every video frame. PCA is applied to reduce the dimensionality of visual features to 30. To synchronize with a video feature, an audio feature is extracted every 1/30 of a second. With 50% overlap, there are 1,470 raw audio samples used to extract into one audio feature, which is a combination of three spectral features and 14 coefficients of MFCC. In our experiments, there are actually 2,311 features extracted after pre-processing. We aggregated features for every 2 seconds as one categorized event. The original DPGMM clustered audio and video features into five and four categories,

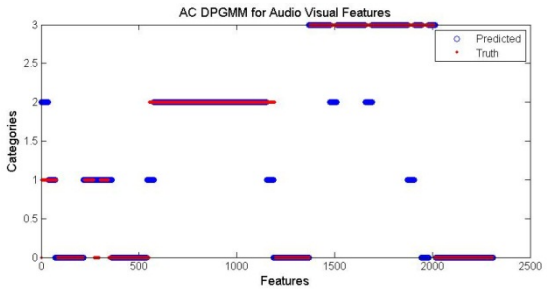

Figure 5: AC-DPGMM for audio-visual features

Table 3. Confusion matrix of 4 activities using audio/video

Acc = 86.07%	B-background, S-speech, C-chair, F-footsteps			
	B	S	C	F
B	785	59	1	10
S	8	134	35	0
C	6	59	576	0
F	36	36	72	494

respectively. The further aggregation step will group audio and video features into four and two categories. The truth categories are manually labeled as: B for background, S for speech, C for moving a chair, F for footsteps while carrying a heavy object.

Figure 3 shows the results of AC-DPGMM for audio features only. Since all categorized events contain sounds, we can cluster them exactly as 4 classes. However, due to the variation of speech features, we get some confusion with the event of moving a chair and the event of footsteps. The total accuracy is about 63.95% and its confusion matrix is shown in Table 1. We can see that there are a lot of footstep features categorized as background, because the sound of a footstep is very short and there is a relatively longer background sound between every two footsteps, and in our labeling process we consider all of them as part of one footstep activity.

Figure 4 shows the results of AC-DPGMM for visual features only. Note that there are actually two categorized events that are clustered, thus observable. One categorized "event" actually contains both real events of moving a chair and footsteps, which cannot be distinguished from each other with video images. Therefore, the accuracy is about 61.75%, lower than the accuracy using the audio only. The confusion matrix in Table 2 shows that all footstep features are categorized into moving a chair.

The combination of audio and visual features using AC-DPGMM shows 4 categories with a total accuracy about 86.07%, as shown in Figure 5 and Table 3. Using both audio and visual information, we can distinguish activities between moving a chair and footsteps. Further, we can detect a person speaking behind a wall whose images cannot be observed via a video camera. Note that the overall accuracy could be much higher if we evaluate the

performance based on events only and with more accurate truth labeling.

We also compare AC-DPGMM with the generative clustering algorithm k-means of four clusters, and the original DPGMM. The result is shown in Table 4. Note that in video only mode, both AC-DPGMM and DPGMM have the same accuracy. This is mainly because only 2 classes are defined during the model learning and those features are clustered consistently in both models. In audio only learning, both Dirichlet models have signification accuracy improvement over k-means. The accuracy of AC-DPGMM using both audio and video has improvement of 23.6% over k-means and 18.8% over DPGMM. Note that the accuracy improvement in combined results using AC-DPGMM is mainly affected by additional aggregation, association and transition steps based on original DPGMM.

Table 4. Comparison of various algorithms.

	k-means (k=4)	DPGMM	AC-DPGMM
Audio	43.01%	63.95%	69.01%
Video	59.58%	61.75%	61.75%
Audio+Video	62.4%	77.22%	86.07%

6. Conclusion

In this paper, we have presented an AC-DPGMM based generative clustering approach for activity recognition using both audio and video at a long range. This model is capable of clustering a set of features as one event with an undefined number of components initially. Association is applied if detections from both sensors are available, and transition is used if detection only happened at one of the sensors. We carried out experiments using audio-video data that can be easily confused using visual sensors only. We compared our results with a generative clustering algorithm based on k-means and the original DPGMM. The combination of both audio and video using AC-DPGMM shows significant improvement in distinguishing some confusing activities in video.

7. Future Work

We note that the performance of clustering results is varied by the selection of feature types and their feature components. In this work, we select audio and visual features and reduced the dimensionality of feature components to an appropriate size based on our experiences. In the future, we would like to learn representative features and their components more intelligently. Also, in the current work, we treat events independently. In real scenarios, one activity may accompany various events in a sequence. We will explore inter-event relations to deal with long and complex activities. We are also interested in using our models to test abnormal activities, such as screaming and gun-shots. Finally a mixture of activities in real-world applications will also be a hard topic to research [2] [3] [5] [14].

8. Acknowledgment

The third author is supported by National Science Foundation (Award # EFRI-1137172). Special thanks to Francine Kergo for proof reading.

References

[1] Cristani, M., Bicego, M. and Murinon, V., "Audio-visual event recognition in surveillance video sequences," *IEEE Transactions on Multimedia,* vol. 9, no. 2, pp. 257-267, 2007.

[2] Dedeoglu, Y., Toreyin, B. U., Gudukbay, U., and Cetin, A. E., "Surveillance using both video and audio," in *P. Maragos et al. (Eds.) Multimodal Processing and Interaction*, Springer, Science+Business Media, LLC, 2008.

[3] Tian, Y., Senior, A. W., Hampapur, A., Brown, L., Shu, C., and Lu, M., "IBM smart surveillance system (S3): event based video surveillance system with an open and extensible framework," in *Machine Vision and Applications*, 2008.

[4] Boiman, O. and Irani, M., "Detecting irregularities in images and in video," *Proc. IEEE International Conference on Comptuer Vision,* pp. 1985-1988, Oct. 15-21, 2005.

[5] Wang, T. and Zhu, Z, "Vision-aided automated vibrometry for remote audio-visual-range sensing," in *Smart Sensor Technologies, eds. K. Iniewsk and M. Syrzycki*, CRC Press, 2013.

[6] Wang, T., Zhu, Z. and Taylor, C. N., "A multimodal temporal panorama approach for movign vehicle detection, reconstruction and classification," *Computer Vision and Image Understanding (CVIU), Speical issue on Advances in Machien Vision Beyong Visible Spectrum,* 2013.

[7] Trabelsi, D., Mohammed, S., Chamroukhi, F., Oukhellou, L. and Amirat, Y., "An unsupervised approach for automatic activity recognition based on hidden markvo model regression," *IEEE Transaction on Automation Science and Engineering,* vol. 10, no. 3, pp. 829-835, July, 2013.

[8] Teh, Y. W., Jordan, M. I., Beal M. J. and Blei, D. M., "Hierarchical dirichlet process," *J. American Statistical Association,* vol. 101(476), pp. 1566-1581, 2006.

Non-Technical Data - Releasable to Foreign Persons.

[9] Khalili, A. H. and Aghajan, Ha., "Multiview actitivy recognition in smart homes with spatio-temporal features," *In Proceedings of the 4th ACM/IEEE ICDSC,* pp. 142-149, 2010.

[10] Krotosky, S. J. & Trivedi, M. M., "Person surveillance using visual and infrared imagery," *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 18, no. 8, 2008.

[11] Kumar, P., Mittal, A., & Kumar, P., "Study of robust and intelligent surveillance in visible and multimodal framework," *Informatic, and International Journal of Informatics and Computing, ACM Society,* vol. 32, no. 1, pp. 63-77, 2008.

[12] Duda, R. O., Hart, P. E., & Stock, D. G., Pattern Classification (2nd Ed.), A Wiley-Interscience Publication, John Wiley & Sons, 2000.

[13] Allen, F. R., Ambikairajah, E., Lovell, N. H. & Celler, B. G., "Classification of a known sequence of motions and postures from accelerometry data using adapted Gaussian mixture models," *Physiol Meas.,* vol. 27(10), pp. 935-951, 2006.

[14] Lin, J. F. S. & Kulic, D., "Automatic human motion segmentation and identification using feature guided hmm for phyiscal rehabiliation exercises," in *In: Robotics for Neurology and Rehabilitation, Workshop at IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2011.

[15] Antoniak, C. E., "Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems," *Annals of Statistics,* pp. 2(6): 1152-1174, 1974.

[16] Fan, W., Bougila, N. & Ziou, D., "Variational learning for finite Dirichlet mixture models and applications," *IEEE Transactions on Neural Networks and Learnign Systems,* vol. 23, no. 5, pp. 762-774, 2012.

[17] Dalal, N. and Triggs, B., "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition*, 2005.

[18] Zheng, F., Zhang, G. and Song, Z., "Comparision of different implementations of MFCC," *J. Computer Science and Technology,* pp. 16(6): 582-589, 2001.