Object-Based On-Line Video Summarization for Internet of Video Things

Shih-Ting Lin[1], Yuan-Hsin Liao[1], Yu Tsao[2], and Shao-Yi Chien[1]
[1]Media IC and System Lab
Graduate Institute of Electronics Engineering and Department of Electrical Engineering
National Taiwan University, Taiwan
[2]Academia Sinica, Taiwan; [3]NTU IOX Center, Taiwan
Email: sychien@ntu.edu.tw

***Abstract*—In order to address the high transmission bandwidth requirement of an Internet-of-Video-Things (IoVT), an object-based on-line video summarization algorithm is proposed to summarize the captured video information at the sensor nodes before being transmitted to the server. It is composed of two stages: intra-view and inter-view stages. In the intra-view stage, human object detector is employed with the proposed human object descriptor. In the inter-view stage, an on-line clustering algorithm with a two-layer K-nearest-neighbor model is also proposed for object clustering. Experimental results show that significant improvement can be achieved when compared with state-of-the-art works.**

I. Introduction

Internet-of-Things (IoT) [1] or machine-to-machine networks (M2M) [2], is the important next wave in the information technology industry. Among different types of IoT, with the rapid development of communication, computation ability and computer vision technologies, Internet-of-Video-Things (IoVT), where video cameras are employed as the major sensors of the IoT, has high potential for wide IoT applications since rich context information can be inferred from video data. The video sensors in an IoVT continuously acquire visual information in real-time and thus generate ultra-big data. How to analyze such a huge quantity of data has become a critical problem.

A key component in an Internet-of-Video-Things is the video sensors. With video sensors, video data of the environment can be analyzed and transmitted to the server. In reality, in order to increase the flexibility and reduce the deployment cost, wireless video sensors are usually preferred. For wireless video sensors, the major design challenges include high transmission bandwidth requirement and high power consumption. From the power analysis of a wireless video sensor [3], the power for data transmission is also one of the major cost of the total system power consumption. Many methods have been studied to reduce the required data transmission bandwidth. Video compression is a technique usually applied to cope with this issue. Nevertheless, the required bandwidth is still large. Another approach which is popular recently is video summarization. By making good summaries for video streams from the network, a great number of redundant data can be removed.

There are many methods for video summarization, such as singular value decomposition [4], sparse coding [5] and graphical model-based solution [6]. These methods all achieve excellent results even on a complex video. However, in order to gain the knowledge of the dataset, most of them are off-line algorithms, which means that all the videos need to be completely recorded before analyzing. Due to the issues of transmission bandwidth, power consumption and storage resource, this kind of methods are not suitable for IoVT. On the other hand, some of the methods only deal with single-view video summarization, which makes them hardly be applied in the network either.

One existing approach that can fit in our scenario is [7], where Gaussian mixture model (GMM) based clustering is employed along with background subtraction for each frame to achieve on-line multi-view summarization. However, if we employ this method to a scene containing multiple and various foreground objects, the results will be disappointed. The key to deal with this problem is the semantic level of data analytics. In this paper, we propose a new object-based method for on-line multi-view video summarization which is able to tackle with complex scenes by increasing the semantic level from frame to object as well as a two-layer K-nearest-neighbor model for object clustering.

The organization of this paper is arranged as follows. In Section II, the proposed method is presented. In Section III, experimental results are shown. Finally, we give the conclusion in Section IV.

II. Proposed Method

Multi-view video streams of a video sensor network can be analyzed from two aspects. We first denote the image captured by the sensor i at time j as $I_{i,j}$. Then, the set

$$E = \{I_{i,j} \mid 1 \leq i \leq V, 1 \leq j \leq t\} \tag{1}$$

represents all the images in the Multi-view video streams within time t. For a set E, we can consider it as the union of subsets Z_v by fixing the index i to v

$$E = \bigcup_{v=1}^{V} Z_v \quad \text{where } Z_v = \{I_{i,j} \mid i = v, 1 \leq j \leq t\} \tag{2}$$

A intuitive explanation for (2) is that an image $I_{i,j}$in the system can be categorized according to the sensor. After the categorization, if we arrange every element in the subset Z_v by

978-1-4673-6853-7/17/$31.00 ©2017 IEEE

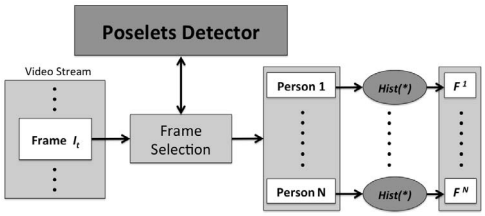

Fig. 1. Block Diagram of Intra-View Stage.

index j, the temporal aspect of the sensor v can be obtained. Similarly, by fixing the index j to u, we can consider the set E as the union of subsets S_u

$$E = \bigcup_{u=1}^{t} S_u \quad \text{where } S_u = \{I_{i,j} \mid 1 \leq i \leq V, j = u\} \tag{3}$$

Arrange all the element in a subset S_u by index i,we then get multiple views of the environment, which is the spatial aspect of the system at time u.

With these two aspects, we divide our system into two consecutive stages, Intra-View and Inter-View Stage. In the former stage, frames are filtered based on temporal aspect. While in the latter stage, distributed view selection will be performed. There are two advantages of our two-stage design for the summarization. First, by splitting the system, the complexity of the task is reduced. Second, filtering frames before Inter-View Stage will shrink the data required to exchange within the network, which decreases the required power consumption.

A. Intra-view Stage: Frame Filtering

In a complex scene, images within a video stream Z_v tend to possess large variance due to a large variety of moving objects, which leads to frustrating results of summarization when utilizing existing algorithms [7]. Since whether an input frame is "unique" in the set Z_v is no longer a good criterion for summarizing, we redefine the important frames as those containing interested objects. This definition perfectly meet our purpose to perform summarization since the information we want to obtain is usually related to the objects-of-interest.

To differentiate important frames from the unimportant ones, we first detect objects in a frame, and categorize them into interested and uninterested ones, which can completely performed in on-line manner. In our algorithm, human objects are taken as the objects-of-interest since human context is usually the most important in applications of IoT. Here, we apply Poselets Detector [8] as our human object detector, since it can not only detect people, but can also localize the torso part, which is beneficial for our algorithm.

The block diagram of the Intra-View Stage is shown in Fig. 1. When a frame I_t is captured by a sensor at time t, human object detection is then performed on I_t with Poselets Detector. If any person n, which is denoted as p^n, is detected, it will be kept and further processed by feature extraction described as follows.

For any p^n, the torso boundary and hip boundary can be obtained from the Poselets Detector. The torso and hip images of p^n, T^n and H^n, can then be cropped from the original frame to represent a person. Apart from the images of torso and hip, the position of p^n in the input frame, $\left(\mathbf{x}^n_{\text{pos}},\ \mathbf{y}^n_{\text{pos}}\right)$, is employed as well. With the three features mentioned above, the entire feature vector for a detected person n is defined as

$$f^n = \left[Hist\left(T^n\right) \quad Hist\left(H^n\right) \quad \left(\mathbf{x}^n_{\text{pos}},\ \mathbf{y}^n_{\text{pos}}\right)\right]^\top, \tag{4}$$

where $Hist\left(*\right)$ represents a function transforms a raw image into a normalized 30-bin color histogram in HSV space, in which H, S and V are all quantized into 10 bins, respectively.

With feature extraction, only the feature information is exchanged between sensors instead of raw image data, which will save the power tremendously.

B. Inter-view Stage: View Selection

After the Intra-View Stage at time t, a portion of images in the set S_t have been abandoned for their redundancy. However, since different sensors of a video sensor network may have overlapped field of views, duplicated information is likely to be transmitted to the server, which will cause a waste of power and storage resource. Therefore, the remaining frames will be further processed by Inter-View Stage to remove unnecessary frames across multiple views in the network.

An intuitive idea of on-line view selection is to calculate an importance score for every view, and only keep the one which has the highest score. Nevertheless, in a complicated scene with a relatively large number of objects, which are people in our task, it is almost impossible for a single senor to completely capture all the objects appear in the network. Consequently, keeping only one particular view will not achieve satisfied result obviously. Since we attempt to filtering out frames without losing any significant events or objects, our approach should be on the basis of objects, which means that it is imperative to perform clustering to identify objects.

According to the result of the Object detection step in Intra-View Stage, a detected people set until time $t-1$ can be denoted as

$$P = \{P_u \mid 1 \leq u \leq t-1\}, \quad \text{where } P_u = \{p^n_{i,j} \mid 1 \leq i \leq V,\ j = u,\ 1 \leq n \leq N\}, \tag{5}$$

with the corresponding features set

$$F = \{F_u \mid 1 \leq u \leq t-1\}, \quad \text{where } F_u = \{f^n_{i,j} \mid 1 \leq i \leq V,\ j = u,\ 1 \leq n \leq N\}, \tag{6}$$

where the newly introduced index n is to represent different human objects. At time t, by exchanging the detected people information between sensors, a new set P_t will be constructed. To exploit the object information effectively during clustering on P_t, our model should preserve the original feature vector of individual person, which is also adapted temporally according to the new input object at each time. Within the various kinds of clustering methods, K-Nearest-Neighbor (KNN) clustering is a solution which possess those features, making it a suitable choice for our scenario. However, another critical problem of multi-view clustering is that the object appearance in every view are usually quite different because of different illumination and camera configurations, which increase the difficulty of clustering. To overcome this obstacle, we propose a new model named as "two-layer KNN model."

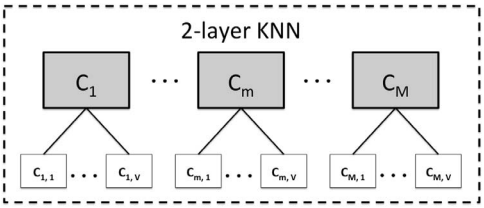

Fig. 2. The structure of two-layer KNN model.

The structure of the two-layer KNN model is shown in Fig. 2. For all cluster $C_m,\ 1 \le m \le M$, we divide it into V sub-clusters

$$C_m = \{C_{m,v} \mid 1 \le v \le V\}, \tag{7}$$

where each sub-cluster corresponds to one of the sensor so that we can avoid the issue of difference in luminance and saturation between sensors. Supposed that until time $t-1$, the set of all feature vectors from view v which have been clustered to cluster m is

$$C'_{m,v} = \{f^n_{v,j} \mid \Phi_c\left(f^n_{v,j}\right) = m \mid 1 \le j \le t-1\} \tag{8}$$

where $f^n_{v,j}$ is the feature vector as shown in (4), and $\Phi_c(*)$ refers to the clustering function that maps a feature vector to its cluster index. We define $C_{m,v}$ as the set of R elements which has largest time index j in $C'_{m,v}$. For every sub-cluster $C_{m,v}$, we also define $\mathrm{CLK}_{m,v}$ as

$$t - \max\{j \mid f^r_{v,j} \in C_{m,v}\} \tag{9}$$

and $\mathrm{POS}_{m,v}$ as

$$f^r_{v,j}(2),\ f^r_{v,j} \in C_{m,v},\ j = \max\{j \mid f^r_{v,j} \in C_{m,v}\}, \tag{10}$$

where $f^r_{i,j}(q)$ refer to the q-th element of $f^r_{i,j}$, and $f^r_{i,j}(2)$ is the position vector as described in (4).

Based on the definitions mentioned above, on-line clustering of detected objects can be performed. The pseudo code is illustrated in Algorithm 1.

Within Algorithm 1, the most important step is computing the distance between $f^k_{i,j}$ and $f^n_{v,t}$. which is define as follows

$$D\left(f^r_{i,j},\ f^n_{v,t}\right) = \alpha\sqrt{\left(D^T_{diff}\right)^2 + \left(D^H_{diff}\right)^2} + (1-\alpha)\,e, \tag{11}$$

where D^T_{diff} and D^H_{diff} is the diffusion distance [9] of torso and hip histogram between $f^r_{i,j}$ and $f^n_{v,t}$ respectively, and e refers to the euclidean distance of $\mathrm{POS}_{m,i}$ and $f^n_{v,t}(2)$. Finally, α is defined as

$$\begin{cases} 1, \text{ if } e \le TH_e \text{ or } \mathrm{CLK}_{m,i} \ge TH_t \text{ or } i \ne v \\ 0.5, \text{ otherwise} \end{cases} \tag{12}$$

where TH_e and TH_t are threshold values.

After clustering for a detected person n by its feature vector $f^n_{v,t}$, the model of two-layer KNN can be updated. We either add the new feature vector to the existing sub-cluster $C_{m,v},$ where $m = \Phi_c\left(f^n_{v,t}\right)$ and abandon the element in $C_{m,v}$ which has smallest time index, or add a new cluster C_{M+1} when the minimal distance between $f^n_{v,t}$ and all features inside the model exceeds some given threshold.

Algorithm 1 On-line Clustering for a Feature Vector $f^n_{v,t}$

Init: Create an empty distance list L_d. Set clustering threshold TH_c.
for all $m \in [1, M]$ **do**
  if $C_{m,v} \ne \varnothing$ **then**
    for all $f^r_{v,j} \in C_{m,v}$ **do**
      Add distance $d^r_m = D\left(f^r_{v,j},\ f^n_{v,t}\right)$ into L_d.
    end for
  else
    for view $i \ne v$ **do**
      for $f^r_{i,j} \in C_{m,i},\ 1 \le r \le R$ **do**
        Add distance $d^r_m = D\left(f^r_{i,j},\ f^n_{v,t}\right)$ into L_d.
      end for
    end for
  end if
end for
$d_{min} = \min L_d$.
if $d_{min} \ge TH_c$ **then**
  return $M+1$.
else
  Create the set L'_d by the smallest K elements from L_d.
  return the index of the cluster which has most elements included in L'_d.
end if

The final step of Inter-View Stage is the View Selection. In order to transmit the most important information of detected objects to the server, we select the most important view, v^*_m, respectively for every cluster C_m. In our system, v^*_m can be obtained as stated below

$$v^*_m = \underset{i,\ \Phi_c\left(f^n_{i,t}\right)=m}{argmax}\ SCORE\left(p^n_{i,t}\right), \tag{13}$$

where $SCORE(*)$ is the score of every detected person provided by the Poselets Detector. Finally, the set

$$\{v^*_m,\ 1 \le m \le M\} \tag{14}$$

will be transmitted to the server as the summarization result of the video sensor network.

III. Experimental Results

In order to evaluate our results of video summarization, we conduct the experiments using three different datasets with eleven videos in total, which are from [10] and our own video sensor network. Example frames of datasets are shown in Fig. 3. The descriptions of our datasets are as follows.

- Intersection 1: We settled three fixed cameras at the road intersection on Fanglan Rd., Da'an Dist near Nation Taiwan University to record a relatively more complex scenario. Multiple interested objects (pedestrian) and uninterested objects (others) appears at the same time frequently.
- Intersection 2: We installed another four cameras at a different road intersection on Fanglan Rd., Da'an Dist from Intersection 1.
- BL-7F: This is the dataset *BL-7F* provided in [10]. The videos inside the dataset were taken in the 7th floor of

Fig. 3. Example frames of the data sets used in the experiments. (a)(b): Intersection 1, (c)(d): Intersection 2, (e)(f): BL-7F

TABLE I. RESULTS OF MULTI-VIEW SUMMARIZATION

Dataset	Method	Summary Length	Precision	Recall	F1 Score
Intersection 1 (3 videos)	GMM[7]	4259	39.4%	27.2%	0.32
	Two-Layer KNN	4533	66.9%	49.2%	0.57
Intersection 2 (4 videos)	GMM[7]	5360	39.6%	29.6%	0.34
	Two-Layer KNN	7378	62.5%	64.4%	0.63
BL-7F (4 videos)	GMM[7]	4680	53.7%	55.0%	0.54
	Two-Layer KNN	5472	56.3%	67.3%	0.61

TABLE II. SUMMARIZATION RECALLS ON OBJECT LEVEL

Dataset	Objects	Methods	
		GMM[7]	Two-Layer KNN
Intersection 1 (3 videos)	person 1	21.7%	59.2%
	person 2	13.5%	49.7%
	person 3	30.7%	32.9%
	person 4	34.1%	40.8%
	person 5	16.0%	91.9%
	person 6	84.6%	1.5%
	person 7	8.4%	76.6%
	person 8	8.5%	77.2%
	Avg.	**27.2%**	**53.6%**
Intersection 2 (4 videos)	person 1	35.7%	71.6%
	person 2	28.5%	73.4%
	person 3	26.9%	75.3%
	person 4	67.1%	91.5%
	person 5	44.7%	73.8%
	Avg.	**40.6%**	**77.1%**
BL-7F (4 videos)	person 1	38.2%	85.5%
	person 2	55.3%	57.5%
	person 3	75.8%	48.7%
	person 4	69.6%	94.6%
	Avg.	**59.7%**	**71.6%**

the BarryLam Building in Nation Taiwan University. The only foreground objects in this dataset are a few people, which make it a relatively more simple dataset.

An on-line multi-view summarization method [7] is implemented as the benchmark, since it has been proven to give state-of-the-art result.

To make the comparison between the results of the two methods more objective, a quantitative metric is used in the experiments. The performance of summarization algorithm is usually evaluated by two main factors: whether the redundant data are removed, and whether the important information are kept. To measure on these two factors, important frames for all the datasets are labeled as the ground truth by human operators who have no knowledge of our work. Then the precision and the recall are calculated, which are related to the first and the second factor mentioned above respectively.

Since our approach abandons frames without objects-of-interest right away during Intra-View Stage and performs view selection for every object respectively, it successfully ignores useless data and keeps important frames in both complex and simple scenes. As shown in Table I, the precision and the recall are much higher than those of GMM [7] method, which implies the better quality of summarization. The F1 score is also calculated and shown in Table I.

To illustrate that our method is able to keep complete information,we ask human operators to mark the important frames for each person as well. The recalls of summarization for each person and the average of object recalls in individual datasets are then evaluated. Our method outperforms GMM [7] remarkably, as shown in Table II.

IV. CONCLUSION

In this paper, we have offered an original approach for on-line multi-view video summarization using object detection and two-layer KNN clustering. From our results, the quality of summarization is considerably improved in compare with existing methods even in a complicated scenes. As a result, our method can be utilized in an Internet-of-Video-Things, or any other video sensor networks set in almost all environments.

ACKNOWLEDGMENTS

This research was supported in part by the Ministry of Science and Technology of Taiwan (MOST 106-2633-E-002-001), National Taiwan University (NTU-106R104045), and Intel Corporation.

REFERENCES

[1] *The Internet of Things ITU Internet Rep.*, 2005 [Online]. Available: http://www.itu.int/internetofthings/

[2] G. Lawton, "Machine-to-machine technology gears up for growth," *IEEE Comput.*, vol. 37, no. 9, pp. 12−15, Sep. 2004.

[3] S.-Y. Chien, T.-Y. Cheng, S.-H. Ou, C.-C. Chiu, C.-H. Lee, V.S. Somayazulu, Y.-K. Chen, "Power consumption analysis for distributed video sensors in machine-to-machine networks," *IEEE J. Emerging and Selected Topics in Circuits and Systems*, vol. 3, no. 1, pp. 55−64. Mar. 2013.

[4] W. Abd-Almageed, "Online, simultaneous shot boundary detection and key frame extraction for sports videos using rank tracing," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2008, pp. 3200−3203.

[5] Y. Cong, J. Yuan, and J. Luo, "Towards scalable summarization of consumer videos via sparse dictionary selection," *IEEE Trans. Multimedia*, vol. 14, no. 1, pp. 66−75, Feb. 2012.

[6] P. Li, Y. Guo, and H. Sun, "Multi-keyframe abstraction from videos," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2011, pp. 2473−2476.

[7] S.-H. Ou, C.-H. Lee, V. S. Somayazulu, Y.-K. Chen, and S.-Y. Chien, "Online multi-view video summarization for wireless video sensor network," *IEEE J. Sel. Topics Signal Process.,* vol. 9, no. 1, pp. 165− 179, Feb. 2015.

[8] L. Bourdev and J. Malik, "Poselets: Body part detectors trained using 3D human pose annotations," in *Proc. Int. Conf. on Computer Vision*, 2009.

[9] H. Ling and K. Okada, "Diffusion distance for histogram comparison," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2006.

[10] S.-H. Ou, C.-H. Lee, V.-S. Somayazulu, Y.-K. Chen, and S.-Y. Chien, "Low complexity on-line video summarization with Gaussian mixture model based clustering," in *Proc. IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, 2014, pp. 1269−1273.