Multi-Source Video Summarization in the Internet of Things: A Social Computing Approach

KLIMIS NTALIANIS[1]

[1]Department of Marketing
Athens University of Applied Sciences
Agiou Spyridonos, Egaleo, Athens
GREECE
kntal@teiath.gr

Abstract: - In this paper a multi-source video summarization scheme is proposed, which is focuses on novel concepts of the Internet of Things. The proposed scheme assumes several different cameras asynchronously recording the same event from different angles, positions, heights etc. One of its main novelties is in the area of content evaluation. More specifically a social computing approach is proposed in order to find out which of the recordings attracts more attention. Attention is estimated by considering a social network environment where each user that records the event can post his/her material and let his/her friends to interact with it. Simulation results are presented, to denote the full potential of the proposed scheme, its advantages as well as important issues for future work.

Key-Words: - Internet of Things, Social Computing, Video Summarization, Multiple Sources

1 Introduction

In the early 2000's, Kevin Ashton was laying the groundwork for what would become the Internet of Things (IoT) at MIT's AutoID lab. Ashton was one of the pioneers who conceived this notion as he searched for ways that Proctor & Gamble could improve its business by linking RFID information to the Internet [1]. The concept was simple but powerful. If all objects in daily life were equipped with identifiers and wireless connectivity, these objects could communicate with each other and be managed by computers. In a 1999 article for the RFID Journal Ashton wrote:

"If we had computers that knew everything there was to know about things—using data they gathered without any help from us—we would be able to track and count everything, and greatly reduce waste, loss and cost. We would know when things needed replacing, repairing or recalling, and whether they were fresh or past their best. We need to empower computers with their own means of gathering information, so they can see, hear and smell the world for themselves, in all its random glory. RFID and sensor technology enable computers to observe, identify and understand the world—without the limitations of human-entered data."

Based on Ashton's "prophesy" the next wave in the era of computing will be outside the realm of the traditional desktop. In the Internet of Things (IoT) paradigm, many of the objects that surround us will be on the network in one form or another. Radio Frequency IDentification (RFID) and sensor network technologies will rise to meet this new challenge, in which information and communication systems are invisibly embedded in the environment around us [2]. In this framework, devices will be interconnected anytime, anywhere on the planet, providing the Internet's advantages in all aspects of daily life [3]. Analysts predict that the IoT will comprise up to 26 billion interconnected devices by 2020, a 30-fold increase from 2009 (www.gartner.com/newsroom/id/2636073). These devices: i) will be able to interact with other devices in an autonomous way with respect to their owners;

ii) will be able to easily crawl the IoT made of billions of devices to discover services and information in a trust-oriented way; and iii) will be able to advertise their presence to provide services to the rest of the network.

But what kinds of services ? Information services, entertainment services, scheduling services etc. These amazingly novel services will bring a new era to the human and non-human beings. This paper focuses on multi-source video summarization, which can be mainly used both in information and entertainment services. The proposed scheme assumes several different cameras asynchronously recording the same event from different angles, positions, heights etc. In order to evaluate the importance of each recording a social computing methodology is proposed. Multi-source video summarization is accomplished by a key-frames extraction algorithm which detects uncorrelated content. Finally, simulation results are presented, to delimit the potential of such applications as well as to set the bases for future work in the field of content summarization in the IoT.

The rest of this paper is organized as follows: in Section 2 state-of-art work in the IoT is presented. Section 3 formulates our problem and provides all necessary definitions. In the Section 4 the proposed multi-source summarization scheme is described, while Section 5 focuses on a simulation analysis. Finally Section 6 concludes this paper.

2 State-of-Art

The IoT, interconnection and communication between everyday objects, enables many applications in many domains [4]. The application domain can be mainly divided into three categories based on their focus [5]: industry, environment, and society. Supply chain management [6], transportation and logistics [7], aerospace, aviation, and automotive are some of the industry focused applications of IoT. Telecommunication, medical technology [8], healthcare, smart building, home [9] and office, media, entertainment, and ticketing are some of the society focused applications of IoT. Agriculture and breeding [10], [11], recycling, disaster alerting, environmental monitoring are some of the environment focused applications. Asin and Gascon [12] listed 54 application domains under twelve categories: smart cities, smart environment, smart water, smart metering, security and emergencies, retail, logistics, industrial control, smart agriculture, smart animal farming, domestic and home automation, and eHealth.

On the other hand, some other specific applications include the work in [13], where traffic lights communicate with and control approaching vehicles upon sensing the presence of pedestrians and bikes. In [14], powered by IoT's ubiquitous identification, sensing, and communication capacities, all objects in the healthcare systems (people, equipment, medicine, etc.) can be tracked and monitored constantly. In [15], in order to prevent and reduce accidents in the mining, the authors propose to use IoT technologies to sense mine disaster signals in order to make early warning, disaster forecasting, and safety improvement of underground production possible. In [16], as more and more physical objects are equipped with bar codes, RFID tags or sensors, the authors propose to the transportation and logistics companies to conduct real-time monitoring of the move of physical objects from an origin to a destination across the entire supply chain including manufacturing, shipping, distribution, and so on. In [17] the authors propose using IoT technologies to construct fire automatic alarming systems in order to raise the firefighting management and emergency management to a new level.

Furthermore there also several other attempts from companies: The Toyota Friend Network is one of the earliest platforms in which data generated by objects, in this case automobiles, is made available in a social network. Developed within the context of a partnership between Toyota and Salesforce, Toyota Friend is a private social network aimed at networking all actors involved in the Toyota car ecosystem, including the cars that become part of the social network as well. Nike+ is another commercial platform in which objects (in this case sensors deployed in basketball shoes) post data in a social network. Nike built around this concept/platform an ecosystem of devices that are sold to customers and services to increase the fidelity of customers to the Nike brand. Higher degrees of autonomy and thus interaction between objects are enabled by the Social Web of Things, which is being developed by scientists at Ericsson Research. The objective is to provide things with more autonomy to help people master the complexity involved in the IoT networking paradigm.

Fig. 1: An overview of the proposed scenario

3 Problem Formulation & Definitions

In order to explain the concept and ideas of the proposed scheme, let us imagine a stadium like the one in Fig. 1, where several spectators watch a baseball match. A number N of the spectators uses a camera (either handheld or head-adjusted like Google glasses, action cams etc.) to record parts or the whole game. Cameras are indicated by Cam 1, Cam 2, …, Cam N. Furthermore, whenever a spectator wants, he/she can post a clip on a social network, so that the clip receives likes/comments or it is shared by friends. To be more precise, let $U = \{1, 2, \ldots, N_U\}$ be the index set of all the users of a social network of N_U users, and thus u_i is the i_{th} user of this social network. Let F_i be the index set of all of the N_F friends of u_i and thus f_{ij} is the j_{th} friend of u_i. Let also I_i be the index set of all of the N_I items posted by u_i and thus i_{ij} is the j_{th} item posted by u_i. The likes, shares and comments for each item posted by each user must be represented and counted. For this reason three vectors are defined:

Denition 1. Let $l_{i,j}$, $p_{i,j}$ and $c_{i,j}$, be respectively the corresponding likes, shares and comments item j, posted from user i, has received. If user i has N_F friends, then:

$$\mathbf{l}_{i,j} = [l_{i,j,f_{i1}}, l_{i,j,f_{i2}}, \ldots, l_{i,j,f_{iN_F}}, l_{i,j,f_{i(N_F+1)}}] \quad (1)$$

$$\mathbf{p}_{i,j} = [p_{i,j,f_{i1}}, p_{i,j,f_{i2}}, \ldots, p_{i,j,f_{iN_F}}, p_{i,j,f_{i(N_F+1)}}] \quad (2)$$

$$\mathbf{c}_{i,j} = [c_{i,j,f_{i1}}, c_{i,j,f_{i2}}, \ldots, c_{i,j,f_{iN_F}}, c_{i,j,f_{i(N_F+1)}}] \quad (3)$$

where $l_{i,j,f_{ik}}$ equals to 1/0 if friend f_{ik} has/has not liked the respective item and similarly $p_{i,j,f_{ik}}$ equals to 1/0 if friend f_{ik} has/has not shared the respective item. At the same time $c_{i,j,f_{ik}}$ equals to the number of comments friend f_{ik} has made to the respective item, while $l_{i,j,f_{i(N_F+1)}}$, $p_{i,j,f_{i(N_F+1)}}$ and $c_{i,j,f_{i(N_F+1)}}$ are used to count respectively the likes, shares and comments the item j has received from everybody else who is not a friend of user i. A slight abuse of notation is already tolerated here, an i as a subscript usually refers to signify a user, where a j as a subscript usually refers to signify a friend f_{ij}, an item i_{ij} etc. relevant to this user. Finally capital Ns are used to signify cardinalities of users (N_U), items (N_I), friends (N_F) etc.

Definition 2. Let us denote as $L_{i,j}$, $P_{i,j}$ and $C_{i,j}$ three scalar variables that count the total number of likes, shares and comments an item j on u_i`s wall has received respectively, as the l_1 norms of their respective vectors (i.e. a summation of their coordinates) as:

$$L_{i,j} = \left\| \mathbf{l}_{ij} \right\|_1 = \sum_{k=1}^{N_F+1} l_{i,j,k} \quad (4)$$

$$P_{i,j} = \left\| \mathbf{p}_{ij} \right\|_1 = \sum_{k=1}^{N_F+1} p_{i,j,k} \quad (5)$$

$$C_{i,j} = \left\| \mathbf{c}_{ij} \right\|_1 = \sum_{k=1}^{N_F+1} c_{i,j,k} \quad (6)$$

Now let us also assume that when a highlight happens in the stadium (e.g. scoring runs, hitting a ball, strikes, player ejection etc.), several of the spectators that record the game, post the recording of the highlight on a social network. This social IoT scenario may raise several questions. However this paper focuses on the following question: can we create an attractive video summary of the highlight by considering multi-source video content as well as social interactions ? Of course our main aims are to optimally cover the highlight and attract as much attention as possible. Towards this direction, in this pilot research, multi-source summarization is accomplished by extracting key-frames from each clip and mixing all key-frames to produce the final summary. The number of key-frames to be extracted from each clip is estimated according to the attention each clip has attracted on social media. For example, if user u_r has posted a clip on his/her wall and the clip has received 20 likes, 15 comments and 5 shares, while user u_t has posted another clip of the same highlight on his/her wall and the clip has received 200 likes, 73 comments and 29 shares, maybe more key-frames should be extracted from the clip of u_t.

4 The Proposed Scheme

In this paper, from each clip key-frames are extracted by minimizing a cross-correlation criterion, so that the selected frames are not similar to each other.

Let us denote by g_k the feature vector of the k_{th} frame of a clip, with $k \in V = \{1, 2, \ldots, N_G\}$ where N_G is the total number of frames of the given clip. Let us also denote by K_G the number of key-frames that should be selected. In order to define a measure of correlation among K_G feature vectors, an index vector is first defined $\mathrm{x} = (x_1, \ldots, x_{KG}) \in W \subset V^{KG}$, where $W = \{(x_1, \ldots, x_{KG}) \in V^{KG}: x_1 < \ldots < x_{KG}\}$ is the subset of V^{KG} containing all sorted index vectors $\boldsymbol{x}$ which contain the frame numbers or time indices of candidate key-frames. Then, the correlation measure among the K_G feature vectors is given by the following equation:

$$R(x) = R(x_1, \ldots, x_{K_G}) = \frac{2}{K_G(K_G-1)} \sum_{i=1}^{K_G-1} \sum_{j=i+1}^{K_G} \rho^2_{x_i, x_j} \quad (7)$$

where ρ_{x_i, x_j} denotes the correlation coefficient between feature vectors $\mathbf{g}_{x_i}$, $\mathbf{g}_{x_j}$, which corresponds to frames with numbers x_i and x_j. Function $R(\boldsymbol{x})$ takes values in the interval [0, 1]. Values close to zero mean that the K_G feature vectors are uncorrelated, while values close to one indicate that the K_G feature vectors are strongly correlated.

Based on the above definition, it is clear that searching for a set of K_G minimally correlated feature vectors is equivalent to searching for an index vector $\boldsymbol{x}$ that minimizes $R(\boldsymbol{x})$. Searching is limited in the subset W, since the correlation measure of the K_G features is independent of the feature arrangement. Consequently, the set of the K_G least-correlated feature vectors is found by:

$$\hat{x} = (\hat{x}_1, \ldots, \hat{x}_{K_G}) = R^{-1}(x) \quad (8)$$

Unfortunately, the complexity of an exhaustive search for obtaining the minimum value of $R(\boldsymbol{x})$ is such that a direct implementation of the method is practically unfeasible. For example, about 264 million combinations of frames should be considered (each of which requires several computations for the estimation of $R(\boldsymbol{x})$) if we wish to select 5 representative frames out of a clip consisting of 128 frames. For this purpose, a logarithmic search algorithm has been proposed in [18] for efficient implementation of the optimization procedure. Although this approach provides very fast convergence, it is highly probable for the solution to be trapped to a local minimum resulting in a sub-optimal solution. In this paper this drawback is alleviated by the use of a guided random search procedure implemented by a genetic algorithm (GA) [19].

4.1 # of Key-Frames per Clip & Complete Summary based on Social Computing

According to the aforementioned analysis, for each specific clip, a number of key-frames is extracted based on correlation. However since the complete summary is an assembly of key-frames from all clips, how many key-frames per clip should

be extracted ? And in which position should be put in the complete summary ?

To answer these questions, in this paper a social computing approach is incorporated. In particular, let us assume that the complete summary should contain a known number of key-frames, say K_C, from n clips. If clip #1 provides K_{G_1} key-frames, clip #2 provides K_{G_2}, etc., then:

$$K_C = K_{G_1} + K_{G_2} + ... + K_{G_n} \quad (9)$$

For simplicity reasons let us assume that clip #1 has received L_1 likes, C_1 comments and it has been shared P_1 times. Similarly clip #2 has received L_2 likes, C_2 comments and it has been shared P_2 times, clip #3 etc., then the number of key-frames per clip can be estimated by:

$$K_{G_i} = \frac{L_i + P_i + C_i}{\sum_{i=1}^{n} L_i + \sum_{i=1}^{n} P_i + \sum_{i=1}^{n} C_i} \cdot K_C, \quad i = 1,...,n \quad (10)$$

Equation (10) assumes that all kinds of interactions between clips and users are of equal importance. Furthermore it is a linear way to estimate the number of key-frames per clip and does not favor the clips that have attracted more attention.

Finally all extracted key-frames are put in order of importance (since it is very difficult to estimate the real chronological order), i.e. the key-frames from the clip that has received the most attention are presented first, while the key-frames from the clip that has received the least attention are presented last. By putting the key-frames in this order, the complete summary is created.

5 Simulation Results

Since currently there are not any IoT standard video datasets for the described or other similar scenarios, the proposed multi-source video summarization scheme has been evaluated by simulation. In particular a multi-source video capturing and social interaction simulator has been created using MATLAB 8.5 (R2015a). The simulator was able: (a) to initialize recordings from several different video sources at different/random time instances, (b) to stop the recording of each video source at a different/random time instance, (c) to virtually "upload" a recorded clip on a virtual "social network" and let simulated users interact with it, (d) to gather interaction values (likes, comments, shares) and (e) to create a feature vector g_k for each frame of a clip.

Source #	Start Time	End Time	Total # of Frames
1	-00:04:11	+00:02:07	9450
2	+00:00:04	+00:01:23	2175
3	-00:00:17	+00:03:45	6050
4	-00:01:11	+00:00:52	3075
5	-00:02:28	+00:01:25	5825
6	+00:00:02	+00:00:57	1475
7	+00:00:14	+00:02:16	3750
8	+00:00:08	+00:04:18	6650

Table I: Details about the eight simulated sources

For visualization purposes an experiment with eight virtual video sources is presented, where their characteristics can be found in Table I. More specifically we assume that the highlight has occurred at time 00(hours):00(minutes):00(seconds). The sign (-) indicates a time before 00:00:00, while the sign (+) indicates a time after 00:00:00. Furthermore we assume usage of the PAL system with 25 frames/sec. As it can be observed the longest clip lasts 378 seconds (9,450 frames) while the shortest clip has a duration equal to 59 seconds (1,475 frames), with an average of about 192 seconds (4,806 frames). This simulation result is not far from reality since people use to continue recording a highlight even after it is completed, depending several times on the reactions of crowds (in case of stadium games).

Source #	L_i	P_i	C_i	K_{G_i}
1	25	2	6	25
2	4	0	1	4
3	17	0	2	14
4	131	6	31	129
5	287	14	37	259
6	12	1	0	10
7	47	1	5	40
8	99	3	23	96

Table II: Interaction values for each clip at iteration #53 and K_{G_i} values for K_C equal to 1.5%.

Now, the total duration of all clips that should be summarized is 1,538 seconds (38,450 frames). Towards this direction and in order to enable the production of an efficient summary, we have let K_C to fluctuate between 1.5 % and 3.5% of the total duration, or between 577 frames and 1,346 frames.

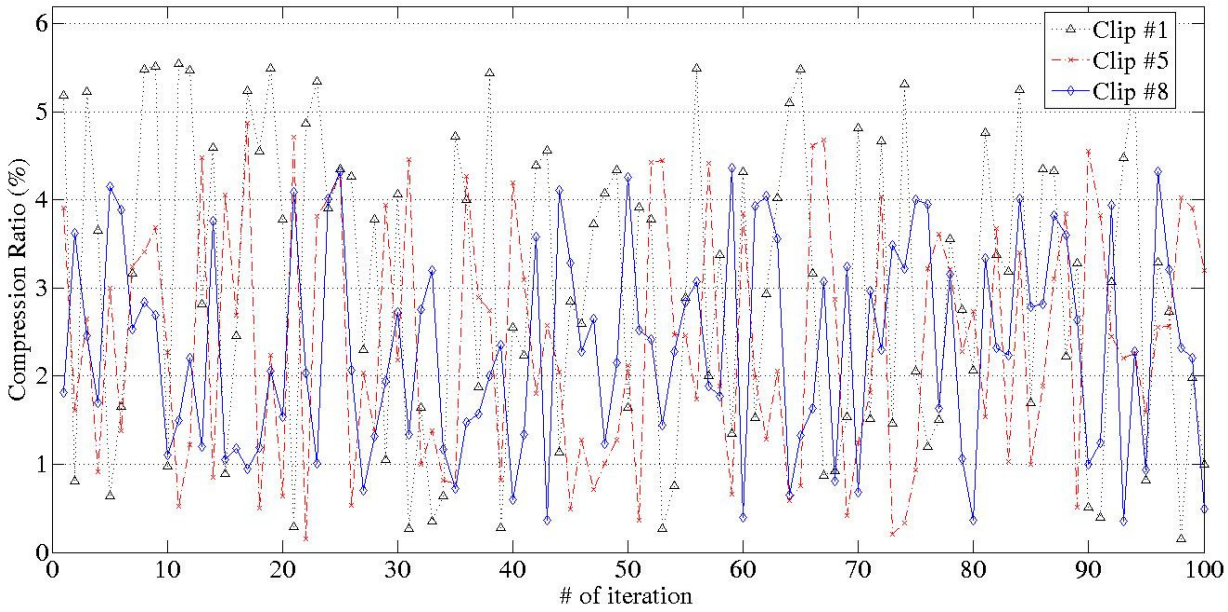

Fig. 2: Percentage of sources #1, 5, 8 that participates in forming the complete summary.

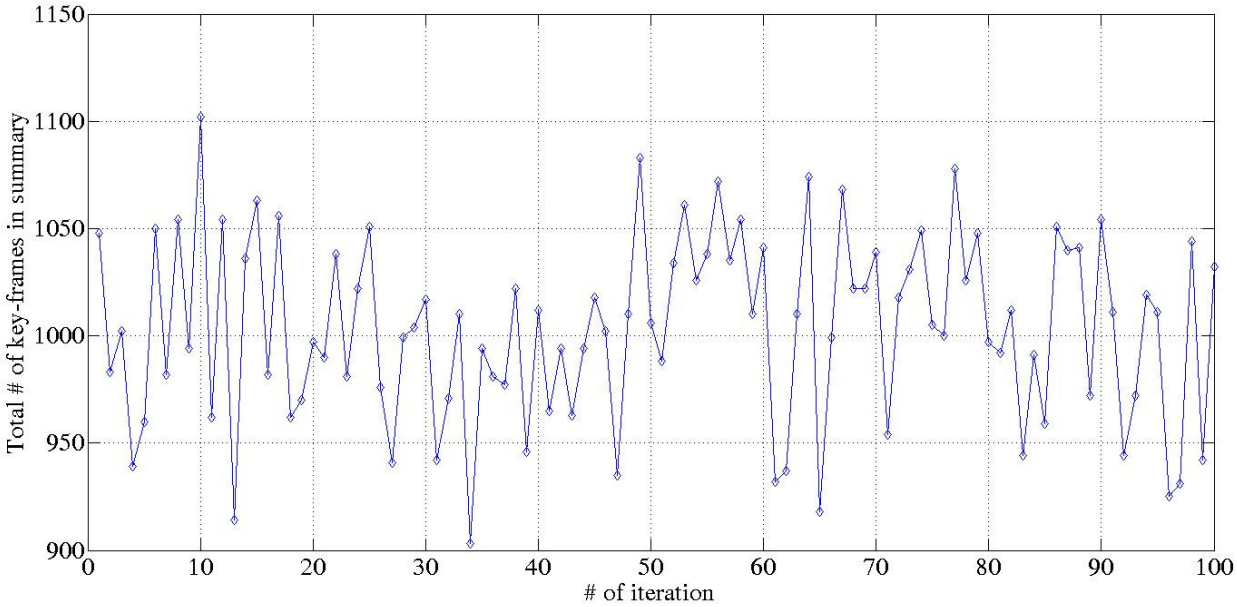

Fig. 3: Total number of key-frames forming the complete summary for 100 iterations.

On average 72 to 168 key-frames should be extracted per source. In order to estimate K_{G_i} (Eq. (10)) for each source, we have also simulated the interactions between each clip and the users of a social network. Since shares, likes and comments are assumed to be of equal importance in the current results, only their total sum (L_i + P_i + C_i) is considered. Additionally the experiment has been repeated several times, providing each time different interaction values for each clip. Table II provides results for iteration #53, where the total number of interactions was 754. In this iteration, the most key-frames (259) are provided by clip #5. Even though K_C is 1.5 %, clip #5 reaches a summarization level of 4.44%, while the less attractive clip #2 reaches a level of only 0.18%. Thus the complete summary is mainly formed by key-frames extracted from the most attractive clip. For visualization purposes, in Figure 2 the fluctuation of the compression ratios (%) for sources #1, 5 and 8 are presented for 100 iterations.

Results have been repeated but this time without setting a value to K_C. Instead, the clip attracting the

most attention provided 3.5% of its frames (key-frames) to the complete summary, while the clip attracting the lowest attention provided 1.75% of its frames. Results are provided in Figure 3 for 100 iterations.

Finally, let us assume that each device records frames of 32 Mpixels size. If frames are uncompressed and 24 bits per pixel are used for encoding, then the total size of the 38,450 frames will be about 3.36 Terabytes. In the experiment of Fig. 3, the maximum number of key-frames in the complete summary is 1,102, while the minimum is 903. Thus the maximum space needed is about 98.5 gigabytes and the minimum is about 80.7 gigabytes, providing an information transmission between 2.35% and 2.87% compared to the initial information.

6 Conclusion

In this paper a multi-source video summarization scheme has been proposed, focusing on the new applications of the Internet of Things. The proposed scheme assumes several different cameras asynchronously recording the same event and performs content evaluation through a social computing approach. Simulation results support the promising concepts introduced by the proposed scheme.

In the future several questions should be considered. What about the number of friends on the social network since different users have different numbers of friends ? What about real time summarization in the IoT ? How complexity issues are going to be solved in case of millions of recordings ? What about aligning (in terms of time) key-frames from different sources, so that users receive a continuous time summary ? These and other issues should be further researched in future works.

References:

[1] http://www.cisco.com/c/dam/en_us/solutions/trends/iot/introduction_to_IoT_november.pdf, November 2013.

[2] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswamia "Internet of Things (IoT): A Vision, Architectural Elements, and Future Directions," *Future Gen. Comput. Syst.*, Vol. 29, No. 7, pp. 1645–1660, 2013.

[3] R. Want, B. N. Schilit, and S. Jenson, "Enabling the Internet of Things," *Computer*, Vol. 1, p.p. 28–35, 2015.

[4] C Perera, A Zaslavsky, P Christen, D Georgakopoulos, "Context aware computing for the internet of things: A survey," *IEEE Communications Surveys & Tutorials*, Vol. 16, No. 1, p.p. 414-454, 2014.

[5] L. Atzori, A. Iera, and G. Morabito, "The internet of things: Asurvey," Comput. Netw., Vol. 54, No. 15, p.p. 2787–2805, Oct. 2010.

[6] L. W. F. Chaves and C. Decker, "A survey on organic smart labels for the internet-of-things," in Networked Sensing Systems (INSS), 2010 Seventh International Conference on, 2010, pp. 161–164.

[7] Y. Chen, J. Guo, and X. Hu, "The research of internet of things' supporting technologies which face the logistics industry," *in Computational Intelligence and Security* (CIS), 2010 International Conference on, 2010, pp. 659–663.

[8] Y.-W. Wang, H.-L. Yu, and Y. Li, "Internet of things technology applied in medical information," in *Consumer Electronics, Communications and Networks (CECNet), 2011 International Conference on*, April 2011, pp. 430 –433.

[9] G. Chong, L. Zhihao, and Y. Yifeng, "The research and implement of smart home system based on internet of things," in *Electronics, Communications and Control (ICECC), 2011 International Conference on*, Sept. 2011, pp. 2944 –2947.

[10] J. Burrell, T. Brooke, and R. Beckwith, "Vineyard computing: sensor networks in agricultural production," *Pervasive Computing, IEEE*, Vol. 3, No. 1, p.p. 38 – 45, Jan.-March 2004.

[11] L. Lin, "Application of the internet of thing in green agricultural products supply chain management," in *Intelligent Computation Technology and Automation (ICICTA), 2011 International Conference on*, vol. 1, 2011, pp. 1022–1025.

[12] A. Asin and D. Gascon, "50 sensor applications for a smarter world," Libelium Comunicaciones Distribuidas, *Tech. Rep.*, 2012.

[13] A. Zanella, N. Bui, A. P. Castellani, L. Vangelista, and M. Zorzi, "Internet of Things for smart cities," *IEEE Internet Things J.*, Vol. 1, No. 1, p.p. 22–32, Feb. 2014.

[14] H. Alemdar and C. Ersoy, "Wireless sensor networks for healthcare: A survey," *Comput. Netw*., Vol. 54, No. 15, p.p. 2688–2710, 2010.
[15] Q. Wei, S. Zhu, and C. Du, "Study on key technologies of internet of things perceiving mine," *Procedia Eng*., Vol. 26, p.p. 2326–2333, 2011.
[16] B. Karakostas, "A DNS architecture for the internet of things: A case study in transport logistics," *Procedia Comput. Sci*., Vol. 19, p.p. 594–601, 2013.
[17] Y. C. Zhang and J. Yu, "A study on the fire IOT development strategy," *Procedia Eng.*, Vol. 52, pp. 314–319, 2013.
[18] Y. Avrithis, A. Doulamis, N. Doulamis, S. Kollias, "A stochastic framework for optimal key frame extraction from MPEG video databases," *Comput. Vision Image Understanding*, Vol. 75, No. 1, p.p. 3-24, July 1999.
[19] N. D. Doulamis, A. D. Doulamis, Y. Avrithis, K. Ntalianis and S. Kollias, "Efficient Summarization of Stereoscopic Video Sequences," *IEEE Trans. on Circuits & Systems for Video Technology*, Vol. 10, No. 4, pp. 501-517, June 2000.