Aggregating Semantic Concepts for Event Representation in Lifelogging[*]

Peng Wang
CLARITY: Centre for Sensor Web Technologies
School of Computing, Dublin City University
Glasnevin, Dublin 9, Ireland
pwang@computing.dcu.ie

Alan F. Smeaton
CLARITY: Centre for Sensor Web Technologies
School of Computing, Dublin City University
Glasnevin, Dublin 9, Ireland
asmeaton@computing.dcu.ie

ABSTRACT
The performance of automatic detection of concepts in image and video data has been improved to a satisfactory level for some generic concepts like indoor, outdoor, faces, etc. on high quality data from broadcast TV or movies. However it remains a challenge to apply this to interpreting the high-level semantics of events as they occur in visual lifelogs from wearable cameras. This is because poorer quality image data and the activities of the wearer make it difficult to automatically categorise them. In this paper, we propose an interestingness-based semantic aggregation and representation algorithm, to tackle the problem of event management and representation in visual lifelogging. Semantic concept interestingness is calculated by fusing image-level concepts which are then exploited to select a representation for the semantic event correlated to various event topics. Experimental results show the efficacy of our algorithm in fusing semantics at the event level, and in selecting representations for event management in visual lifelogging.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; H.5.0 [**Information Interfaces and Presentation**]: General

General Terms
Algorithms, Experimentation, Measurement

Keywords
visual lifelogging, concept interestingness, concept aggregation, keyframe selection, semantic fusion

1. INTRODUCTION
It is now quite practical for researchers to investigate the underlying patterns of our daily lives following the widespread availability of lightweight devices such as mobile phones endowed with sensing abilities through built-in cameras and other sensors. The vision of using technology to automatically record everything that happens to us is called lifelogging [2]. Steve Mann is a pioneer who tried to capture what he saw through video cameras mounted on his head [9]. Following Steve Mann, there has been much research work on life assistance using visual lifelogging techniques. Microsoft Research in Cambridge have developed the SenseCam to capture everyday lives and there is evidence that these images can improve peoples' memory abilities [11]. At MIT, an experiment was carried out using Bluetooth-enabled mobile telephones to measure information context in order to identify the deep social patterns in user activities [7]. In [17], Vemuri and Bender presented a memory re-finding use of lifelogging which is called "iRemember". In their research they recorded audio clips as the main information to navigate memory.

As a new form of multimedia data, the management of visual lifelogs should involve semantic indexing and retrieval for which much preliminary work has already been done in other domains on bridging the gap between low-level features (colour, texture, shape, etc.) and high-level semantics (concepts, topics, etc.). State-of-the-art techniques use statistical approaches to map low level features to concepts which are then fused to relate to high level semantic topics [15]. According to the TRECVid benchmark [14], acceptable results for concept detection have been achieved already in many cases particularly for parts of concepts and related tasks when there exists enough annotated training data. A large-scale concept ontology (LSCOM) has been developed for standardizing multimedia semantics in the broadcast TV news domain [10]. As a framework, the LSCOM effort also produced a set of use cases and queries along with a large annotated data set of broadcast news video. The individual concepts detected by standalone classifiers are usually fused for topic-related filtering which demands a high level of classification accuracy [3]. However, in the visual lifelogging domain, the burden of improving detection accuracy is more severe given the visual diversity of visual lifelog content and the large variety of concepts compared to, say, broadcast TV news. Even the images captured passively within the same lifelogged event might have significant perceptual differences due to wearer's movement. An approach to fuse detected semantic concepts in terms of *event-level* topics is demanded by aggregating useful concepts in a stable manner. In this paper, an interestingness-based concept aggregation algo-

---

rithm is proposed to extract concepts under the consistency of the semantics of the lifelog *event*. Concept aggregation refers to combining and ranking concepts detected at the image level to obtain the most unique and specific concepts at the event level.

The rest of the paper is organized as follows: in Section 2, we describe the construction of a semantic concept space in our passive-capture visual lifelogging domain. Our algorithm for interestingness-based concept aggregation is discussed in Section 3, followed by the semantic selection of event representations in Section 4. The experimental setup and results analysis is presented in Section 5. Finally, we close the paper with conclusions.

2. CONSTRUCTING A CONCEPT SPACE
In the lifelogging domain, visual concepts are important semantic sources for interpreting everyday events which are the subject of the lifelog. In order to record what we see and what we do unobtrusively, we employed SenseCam (shown in Figure 1) as a wearable device to log details of users' lives. SenseCam is a lightweight passive camera with several sensors built-in which captures the view of the wearer with its fisheye lens. By default, images are taken at the rate of about one every 50 seconds while the onboard sensors can help to trigger the capture of pictures when sudden changes are detected in the environment of the wearer. Some typical example events from SenseCam data are shown in Figure 2.

Figure 1: SenseCam (right as worn by a user).

Semantic concepts appearing in SenseCam images can be used to construct a concept space which is defined as a linear space with a set of concepts as the base vectors. In order to reflect the semantics for events, every concept representing any event should represent one of the dimensions in semantic space, and the projection of an event onto concepts is the co-occurrence information among the concepts. However, different concepts have different impacts on event interpretation. Concepts which are neither too general nor too specific should be selected in the semantic space to reduce dimensionality and noise (where noise refers to erroneous classification in this paper) for concept detection, similar to the way index terms are chosen to represent documents in information retrieval. This means we should include event-centered concepts with decent frequency, and exclude general and over-specific ones.

In order to ensure high coverage of this space, we elaborate the selection of a set of concept bases according to the generalization of entities in the semantic space. During the procedure to determine the target concepts, a subset of Sense-Cam images were first inspected to determine the typical concepts employed among the LSCOM [10] and MediaMill

concept ontologies [16] some of which might have been applicable to our domain. As expected, we found that some of the LSCOM concepts, for example weapon, government leader, etc., are never useful or even encountered in the lifelogging domain so while the hierarchical structure of LSCOM might have been useful, the actual concepts were not. We also considered including some concepts beyond the ones in these ontologies which have high frequency of appearance among SenseCam images. The generalisability of each concept is thus investigated across collections and users as a criterion to refine the concept set iteratively [2]. We used a final set of 27 concepts in constructing the concept space in our SenseCam-based event interpretation. These 27 concepts are shown in Table 1 as a universal set organized into general categories of objects, scene/setting/site, people and events. Note that the methodology in this paper is generic and can be extended to larger concept sets as well.

Table 1: SenseCam concept sets

Objects	screen, steering wheel, car/bus/vehicles
Scene/Settings/Site	indoor, outdoor, office, toilet/bathroom, door, buildings, vegetation, road, sky, tree, grass, inside vehicle, view horizon, stair
People	face, people, hand
Event	reading, holding cup, holding phone, presentation, meeting, eating, shopping

Following the state-of-art in concept detection, we employed the popular generic SVM learning algorithm for concept detection. Two MPEG-7 features were extracted for each image, Scalable Colour (12 bins) and Colour Layout (64 bins) forming 76-dimensional feature vectors. For the results presented in this paper, SVM-Light [8] was employed with the radial basis function (RBF) as a kernel, $K(\mathbf{a}, \mathbf{b}) = exp(-\gamma \|\mathbf{a} - \mathbf{b}\|^2)$. The parameter settings were determined through iterative searching among parameter combinations. Classification models were trained for different concepts yielding a 27-dimensional confidence vector for each image. There are disadvantages of having concepts work independently and in isolation as in our 27 but how to overcome these and have concept detection work as a group of concepts which learn from each other during the detection phase, is currently not a well-established technique in machine learning and so not used in concept detection.

3. INTERESTINGNESS-BASED CONCEPT AGGREGATION
An interestingness-based concept aggregation method is proposed which fuses the occurrence of concepts at the image level in order to reflect semantic consistency within the same event, as well as differences among individual events. It is important to realise that a single lifelog event such as sitting on a bus, walking to a restaurant, eating a meal, watching TV, etc. consists of many, usually hundreds, of individual SenseCam images. In the case of sitting on a bus, where there is little movement by the wearer, most SenseCam images are the same whereas giving a lecture, for example, where the wearer is moving around, generates a large range of dissimilar images. How we construct representative concepts for events is now described.

3.1 Event Concept Interestingness

In visual lifelogging, successively captured images may have quite different visual appearance and a variety of concepts detected, unlike traditional video for which two successive frames within the one shot will be visually very similar. This makes it impossible to use the concepts from one single lifelogged image to infer the semantics of a whole event. Meanwhile, different concepts play different roles in interpreting event topics. For example, in analyzing concepts for a 'meeting' event, we can detect such concepts as 'indoor', 'office' and 'face'. As 'indoor' is not a unique concept for 'meeting' compared to other events such as 'working', 'shopping' that also have the concept 'indoor' occurring, it should be ranked lower while concepts like 'office' and 'face' are better representations for 'meeting'. The interestingness-based concept aggregation is motivated by the notion that the best descriptive concepts for an event should be the most unique across the collection yet representative, in order to differentiate the given event from others; meanwhile the concept should also have relatively high frequency within the event. This is the same rationale as $tf \times IDF$ weighting in standard information retrieval.

To simplify the problem domain, we limit event coverage to within the range of a day because most of the time users interpret events within a daily basis. The algorithm could easily be extended to a week or month basis which has broader time intervals. To formalize the calculation, we assume a universe of concepts C. Let $\{E_1, E_2 \ldots E_N\}$ be the sequence of events in a given day. Event E_i is represented by successive images $I^{(i)} = \{Im_1^{(i)}, Im_2^{(i)} ... Im_m^{(i)}\}$. Each image Im_j^i might have several concepts detected, we assume the concepts appearing in image $Im_j^{(i)}$ are $C_j^{(i)} = \{c_{j1}^{(i)}, c_{j2}^{(i)} ... c_{jn}^{(i)}\}$. Then the frequency of concept c occuring in event E_i is calculated in the form of $f(c, E_i) = \sum_{1 \leq j \leq m} 1\{c \in C_j^{(i)}\}$, where $1\{\cdot\}$ is the indicator function.

The weight for each concept $c \in C$ for E_i given the above assumption is:

$$w(c, E_i) = \frac{f(c, E_i)}{\sum_{1 \leq j \leq N} f(c, E_j) + \xi} \qquad (1)$$

The definition above can satisfy the assumptions in [6] as follows:

1) Frequently occurring concepts show semantic consistency within the event and should be selected as candidates.

2) Concepts appearing more during E_i than during other events are more unique and should have higher weights.

Concepts detected at the image level are prone to noise and suffer from misclassification. ξ in the denominator of (1) is used to filter misclassified concepts with very low frequency. However, the aggregation at the event level can filter misclassified concepts and only consistent concepts having a higher weight will be selected.

3.2 Semantic Aggregation of Concepts

In the event segmentation stage, each event is separated from others using sensor readings from the SenseCam's onboard sensors [5], and a keyframe is selected as the best representative image for each event [4]. Though concept detection is easily affected by noise at the image level, our concept aggregation fuses the dominant concepts from the event level which shows greater robustness to concept detection noise. The fusion procedure returns the Top-k concepts for event E_i ranked according to concept interestingness as $\{c_1^{(i)}, c_2^{(i)} ... c_k^{(i)}\}$, where interestingness weight $w(c_j^{(i)}, E_i) \geq w(c_{j+1}^{(i)}, E_i)$. The choice of Top-$k$ value can be modified, which will be explored in the experiments in Section 5.

Keyframes	Event details	Event concepts	Aggregated concepts
Event_1		Indoor	People
		Office	Indoor
		People	Office
		Hands	Hands
		Screen	Face
		Face	Screen
		Meeting	Reading
		Reading	Meeting
Event_2		Indoor	Outdoor
		Outdoor	Buildings
		Buildings	Sky
		Sky	Tree
		Office	Vegetation
		People	Road
		Tree	Grass
		Vegetation	People

Figure 2: Event-level concept aggregation.

The main contribution of this paper by applying concept aggregation is representing event with a vector of concepts which not only reflects event semantics, but also facilitates event visual representation, i.e. keyframe selection. Some examples are shown in Figure 2 in which the resulting concepts from the aggregation algorithm are listed. Due to the disadvantages of the single concept classifier, only those concepts with high confidence can be regarded as true from each image. Thus some concepts which might be more relevant at the event level are easily missed. In Figure 2 we can see that the keyframe may be visually representative of the event but we are not sure if it is semantically representative. In Event_1, only two concepts can be detected from keyframe, namely 'indoor' and 'office', forming a concept vector $C_{kf1} = \{indoor, office\}$. From these two concepts there is ambiguity as to the nature of Event_1. The aggregated method ranks the more unique concepts higher through intestingness weight vector $\mathbf{v_{e1}} = (0.037, 0.034, 0.022, 0.020, 0.011, ...)$, of which each value represents the weight for 'people', 'indoor', 'office', 'hands', 'face' and so on. These concepts have more correlation with event semantics such as 'talking' ('people', 'face') and 'using computer' ('hands','screen'). These two types of activities reflect the core semantics of Event_1.

4. EVENT SEMANTIC REPRESENTATION: A VSM-LIKE PARADIGM

As a widely used search model, the Vector Space Model (VSM) [1] is known as one of the most typical models in information retrieval. In VSM, all entities including documents, queries and terms, are represented as vectors [12]. Using term vectors as the basis in vector space, both document and query vectors are built as linear conmbinations of

the term vectors. The evaluation is then done by analyzing the correlation between the vectors as the relationship between query and document. In this section, we employ the VSM model as the representation for events.

4.1 Semantic Vector Similarity

To quantify the relationship between entities in the semantic space, we will discuss the similarity of concept lists. The $tf \times IDF$ weight is used as the most efficient weighting definition in Vector Space Model where both documents and queries are associated with t-dimensional vectors $\mathbf{v_j} = (w_{1j}, w_{2j}, \ldots, w_{tj})$, where each dimension is a weight and t is the size of lexicon. Traditional vector similarity measures can be employed to quantify the relevance between two vectors, such as inner product $(\mathbf{v_i} \bullet \mathbf{v_j})$ or cosine of the angle among those two vectors as $(\mathbf{v_i} \bullet \mathbf{v_j})/(||\mathbf{v_i}|| \times ||\mathbf{v_j}||)$.

However, the semantic contribution of each dimension to the vector is ignored by these measures. Especially, it worsens the case if terms, which are concepts in image or video retrieval, can not be detected perfectly. The noise introduced by imperfect concept detection will degrade the performance. For example, assume we have three semantic vectors: $\mathbf{v_1} = (0.1, 0.2, 0.1)$, $\mathbf{v_2} = (0, 0.2, 0)$, $\mathbf{v_3} = (0.2, 0.1, 0.2)$, whose components represent the weight for different concepts representatively. Though cosine similarity $sim(\mathbf{v_1}, \mathbf{v_2})$ is equal to $sim(\mathbf{v_1}, \mathbf{v_3})$, we prefer $\mathbf{v_2}$ to approach $\mathbf{v_1}$ because they semantically emphasize the same concept. Besides, the low weights in $\mathbf{v_1}$ such as 0.1 are more likely to be affected by noise introduced by concept detection, making the similarity unstable.

With this motivation, we define the similarity which considers both set agreement and rank consistency of two concept vectors and apply the measurement in judicious selection of an event keyframe. The similarity is shown as the following equation:

$$sim(C_i, C_j) = \frac{1}{|C_i \bigcup C_j|} \sum_{k=1}^{|C_i|} \sum_{l=1}^{|C_j|} \frac{1\{C_{ik} = C_{jl}\}}{abs(k-l)+1} \quad (2)$$

where C_i, C_j stands for two concept vectors aggregated by approaches described in Section 3.2, $|C_i \bigcup C_j|$ is the cardinality of the set consisting of the union of two concept sets. $abs(k-l)$ gives the absolute value of ranking difference for the same concept in two vectors. The added "1" in the denominator is used to avoid division by zero.

The concept vectors are regarded as high-level features for interpreting event semantics. To demonstrate the similarity for high level features, let's revisit the examples in Figure 2. We choose Top-5 concept vector for Event_1 for simplicity, which are $C_{e1} = \{people, indoor, office, hands, face\}$. According to the definition above, the similarity of C_{e1} and C_{kf1} for the keyframe is 0.2 for Event_1. With the same manner, the semantic similarity between keyframe ($C_{kf2} = \{indoor\}$) and event for Event_2 is 0.028. Event_2 has much lower vector similarity due to the existence of sub-events with disjoint semantics of 'outdoor' and 'indoor'.

4.2 Semantics-based Keyframe Selection

Up to 3,000 images are captured on a typical day using SenseCam. Without an effective indexing mechanism, looking through these one-by-one is not a scalable approach to navigating such a collection. The event-centric media representation approach we propose here should yield high semantic consistency with regard to representing the high-level meaning of lifelog events. We employed aggregated concepts rather than low-level features, aiming to select a keyframe for an event which is most relevant to the *whole event* semantics.

As described above, event semantics are represented in the form of high-level features by a concept vector within which concepts are ranked according to uniqueness. Assuming that event $e = s_1, s_2...s_N$ has the concept vector C_e, each image s_i has concept vector C_i. Both C_e and C_i are ranked in terms of the methodology in Section 3.2. The keyframe is chosen as satisfying:

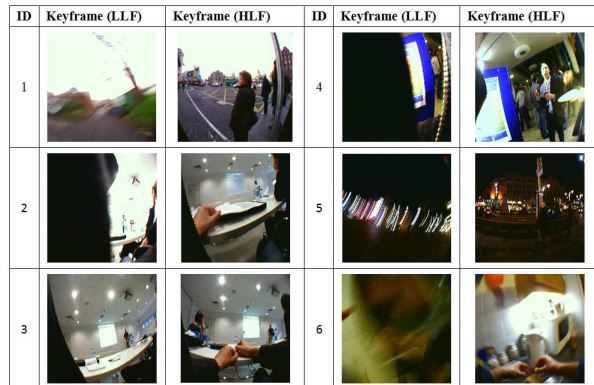$$s^* = argmax_{s_i \in e, 1 \leq i \leq N} sim(C_i, C_e) \quad (3)$$

ID	Keyframe (LLF)	Keyframe (HLF)	ID	Keyframe (LLF)	Keyframe (HLF)
1			4		
2			5		
3			6		

Figure 3: Semantic representation for events.

To illustrate the advantages of this approach, Figure 3 demonstrates examples from which the keyframes using low-level features (LLFs) and high-level features (HLFs) are compared. Six events are randomly selected from one day. The representations selected by high-level features have obviously better image quality than the ones selected based on low-level features, especially for events 1, 5 and 6. Objects are hardly recognizable in the LLF representation for event 1 and 6 due to motion blur. The images with higher quality often have more detail and concept information, so they are naturally selected as better representations using HLFs. In events 2, 3 and 4, the HLF representations are better than the LLF ones because of wider visual fields. Even during darkness, the HLF selection approach will choose images with more detail and better quality as shown for event 5.

5. EXPERIMENT AND EVALUATION

As mentioned earlier in this paper, SenseCam images have very different visual characteristics to the video keyframes used in the TRECVid benchmark [13, 14] and so we could not evaluate the performance of our concept detection on the TRECVid datasets. Thus an experiment was carried out on 6 participants' SenseCam image logs. The participants are all researchers in our lab and have been wearing SenseCam for varying lengths of time. The effect of

interestingness-based semantic keyframe selection is compared with the baseline which is the selection of the middle image as a representation for an event, the same technique as is used for keyframe selection in video. Details of the data are shown in Table 2 indicating a total of 1,055 events composed of 96,217 individual images.

Table 2: Experimental data set

Users	User1	User2	User3	User4	User5	User6
Events	300	248	242	168	70	27
Images	26,062	25,341	19,233	18,085	6,097	1,399

Concepts were first detected at image level, followed by interestingness-based aggregation to model event semantics. We empirically choose the value $\xi = 200$ in Equation 1 considering the fact that most events have less than 200 images. Image-event semantic similarities are calculated to select the most similar image to the event semantics. In [4], a fusion of the Contrast and Saliency Measures in exploiting image quality show promising user judgement scores, which are no less satisfactory than more complicated fusions taking Colour Variance, Global Sharpness or Noise Measure into account. We employ the Contrast Measure and Saliency Measure from [4] as two measures to evaluate resulting keyframe quality. The Contrast and Saliency scores are calculated and normalized on a Max-Min scale respectively. To decrease the effect of external factors such as life patterns of individuals and characteristics of different SenseCam lenses, we analyze the results of our algorithm on a per-user basis.

Our semantic similarity measurement is tested on resulting Contrast and Saliency scores. Figure 4 shows the Contrast difference of selected keyframes by semantic similarity (SS) defined as Equation 2 and by cosine similarity (COS) on one random user's dataset. The averaged Contrast scores over all event numbers are 0.477 and 0.459 using SS and COS measures respectively. From Figure 4, it's obvious that keyframes selected by SS measure have better Contrast quality. The same happens for the Saliency measure as shown by Figure 5, where averaged Saliency scores using the SS measure outperforms the COS measure by 15%. The semantic similarity also shows significant advantages over other measures like inner product, Euclidean and so on but we will not elaborate the details here.
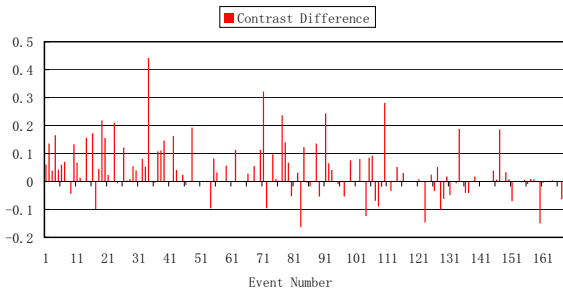

Figure 4: Contrast difference (SS-COS).

In Figure 6, the improvement on average values of the Contrast and Saliency Measures with semantics-based representation are shown for each user. Both measurements are significantly enhanced over the baseline for all participants.
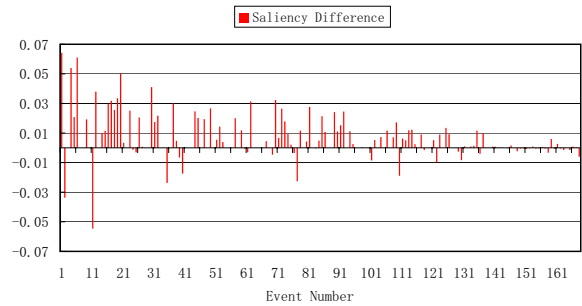

Figure 5: Saliency difference (SS-COS).

Note that user5 is using an old SenseCam whose lens is blurred yet the semantics-based algorithm still performs well showing the robustness of our semantic modeling.
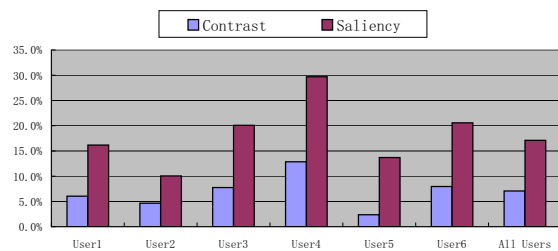

Figure 6: Contrast vs. Saliency Measure.

Modeling complexity is modified in our experiments by changing the selection of Top-k ranking of concept vectors to test the effect of event semantics on the selection of representative images. Figure 7 shows the dependence of keyframe quality on the semantics of events, by selecting the Top-k concepts. Results are depicted using an equally-weighted image quality value of Contrast (0.5) and Salience (0.5). For illustration, we randomly selected three participants' fused image quality scores and compared with their corresponding baseline values. With parameter k decreasing, the fused quality of semantics-based representation drops after k is less than 10. The correlation of quality score with choice of k demonstrates the impact of semantics of events on keyframe selection. When just a little semantics are employed, see $k \leq 2$, the quality score curves intersect with their own baselines, showing no obvious improvement. This also shows that our similarity measure is appropriate in deciding the relationship for concept-based semantics.
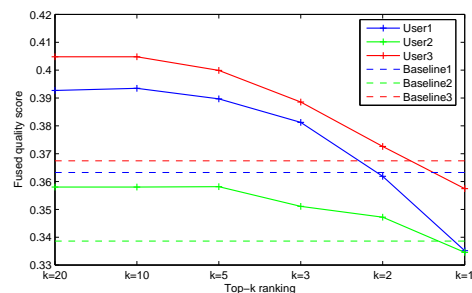

Figure 7: Correlation of quality with Top-k.

Figure 8 compares the number of concepts detected from each selected keyframe. When more concepts are used, e.g. $k = 20$ or 10, keyframes tend to contain more semantics about the events (nearly half have 3 or 4 concepts). Similar to image quality in Figure 7, the number of concepts in the representation decreases with smaller k values. Meanwhile, the representativeness of keyframes drops and less detail about the represented events are found. When only the first concept is selected from the event concept vector, say $k = 1$, the semantics reflected in the semantics-based representation is almost similar to the baseline.
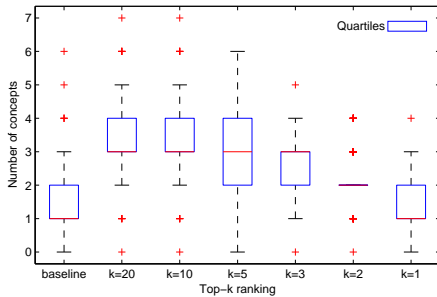

Figure 8: Concept number in single representation.

As demonstrated above, the image quality and potential concepts from the keyframe selected based on semantics shows strong correlation with the choice of k. When more semantic information is applied ($k \geq 5$), our algorithm performs well in selecting keyframes which are more representative and of better quality. Our interestingness-based event aggregation not only reflects semantics of events but also provides a computable platform in comparing semantic relationships such as similarity in the same concept space.

6. CONCLUSIONS

An interestingness-based aggregation algorithm is proposed to deal with the issue of event-level semantic fusion in visual lifelogs. The approach is shown to be effective and robust both in representing event semantics and selecting keyframes for events. Experimental results demonstrate that the semantically selected keyframes have better properties than a baseline method in many aspects such as image quality and concepts appearing as well as semantic similarity with events. Our future work is to employ this algorithm in mining semantic relationships between events, complementary to low-level feature-based similarity measurements, to interpret event patterns in lifelogging.

7. REFERENCES

[1] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley, 1st edition, May 1999.

[2] D. Byrne, A. Doherty, C. G. M. Snoek, G. Jones, and A. Smeaton. Everyday concept detection in visual lifelogs: validation, relationships and trends. *Multimedia Tools Appl.*, 49(1):119–144, 2010.

[3] M. Christel, M. R. Naphade, A. Natsev, and J. Tesic. Assessing the filtering and browsing utility of automatic semantic concepts for multimedia retrieval. In *Proc. of the 2006 Conf. on Computer Vision and Pattern Recognition Workshop*, page 117, Washington, DC, USA, 2006. IEEE Computer Society.

[4] A. Doherty, D. Byrne, A. Smeaton, G. Jones, and M. Hughes. Investigating keyframe selection methods in the novel domain of passively captured visual lifelogs. In *Proc. of the 2008 International Conf. on Content-based Image and Video Retrieval*, pages 259–268. ACM, 2008.

[5] A. Doherty and A. Smeaton. Automatically segmenting lifelog data into events. *Int. Workshop on Image Analysis for Multimedia Interactive Services*, 0:20–23, 2008.

[6] M. Dubinko, R. Kumar, J. Magnani, J. Novak, P. Raghavan, and A. Tomkins. Visualizing tags over time. In *Proc. of the 15th International Conf. on WWW*, pages 193–202. ACM, 2006.

[7] N. Eagle and A. Pentland. Reality mining: sensing complex social systems. *Personal Ubi. Comput.*, 10(4):255–268, 2006.

[8] T. Joachims. *Making large-scale support vector machine learning practical*, pages 169–184. MIT Press, Cambridge, MA, USA, 1999.

[9] S. Mann. Wearable computing: A first step toward personal imaging. *Computer*, 30(2):25–32, 1997.

[10] M. Naphade, J. Smith, J. Tesic, S.-F. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis. Large-scale concept ontology for multimedia. *IEEE Multimedia*, 13(3):86–91, July 2006.

[11] A. Sellen, A. Fogg, M. Aitken, S. Hodges, C. Rother, and K. Wood. Do life-logging technologies support memory for the past? An experimental study using SenseCam. In *Proc. CHI 2007*, pages 81–90, New York, NY, USA, 2007.

[12] I. R. Silva, J. N. Souza, and K. S. Santos. Dependence among terms in vector space model. In *Proceedings of the International Database Engineering and Applications Symposium*, pages 97–102, Washington, DC, USA, 2004. IEEE Computer Society.

[13] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and trecvid. In *MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, pages 321–330, New York, NY, USA, 2006. ACM Press.

[14] A. F. Smeaton, P. Over, and W. Kraaij. High-Level Feature Detection from Video in TRECVid: a 5-Year Retrospective of Achievements. In A. Divakaran, editor, *Multimedia Content Analysis, Theory and Applications*, pages 151–174. Springer Verlag, Berlin, 2009.

[15] C. Snoek, B. Huurnink, L. Hollink, M. de Rijke, G. Schreiber, and M. Worring. Adding semantics to detectors for video retrieval. *IEEE Trans. on Multimedia*, 9(5):975–986, 2007.

[16] C. Snoek, M. Worring, J. C. Van Gemert, J.-M. Geusebroek, and A. Smeulders. The challenge problem for automated detection of 101 semantic concepts in multimedia. In *Proc. of the 14th Annual ACM International Conf. on Multimedia*, pages 421–430, New York, NY, USA, 2006. ACM.

[17] S. Vemuri and W. Bender. Next-generation personal memory aids. *BT Technology Journal*, 22(4):125–138, 2004.