ONSUM: A System for Generating Online Navigation Summaries

Yogesh Girdhar and Gregory Dudek

Abstract— **We propose an algorithm for generating navigation summaries. Navigation summaries are a specialization of video summaries, where the focus is on video collected by a mobile robot, on a specified trajectory. We are interested in finding a few images that epitomize the visual experience of a robot as it traverse a terrain. This paper presents a novel approach to generating summaries in form of a set of images, where the decision to include the image in the summary set is made online. Our focus is on the case where the number of observations is infinite or unknown, but the size of the desired summary is known. Our strategy is to consider the images in the summary set as the prior hypothesis of the appearance of the world, and then use the idea of Bayesian Surprise to compute the novelty of an observed image. If the novelty is above a threshold, then we accept the image. We discuss different criterion for setting this threshold. Online nature of our approach allows for several interesting applications such as coral reef inspection, surveying, and surveillance.**

I. INTRODUCTION

In this paper, we present a technique by which a robot, equipped with a camera, can continuously monitor a location, and at any requested time, present us a summary consisting of a small set of with visually notable images which summarize its visual experience. Such image set is called a *Navigation Summary*. In this paper, we are interested in the online version of the problem where the decision: whether or not to include an image in the summary set, is made immediately after it has been observed. This online constraint allows the system to be used to trigger some physical event, and hence is useful for several different applications such as surveillance, monitoring of patients, and sensor placements.

It is easy for a robot to record all of its observations. However, eventually a human must look at all the data and take appropriate actions. For a time critical task such surveillance a robot which can present not only a summary of its visual experience at any given time, but also alert us on any surprising observations, can be of great help.

While navigation summaries sometimes are based on the use of GPS data as well as pure video information, this paper is concerned with the acquisition and use of video data alone for the online summarization process. Recall that *online algorithms* refer the class of methods that produce results incrementally as data is received, as opposed to *offline* or batch processing that waits until all the data is collected.

The problem of identifying summary images is related to the problem of identifying landmark views in a view based mapping system. A good example is work on View-based maps by Konolige et. al. [10]. In this work, the goal is to identify a set of representative views and the spacial constraints among these views. These views are then used to localize the robot. With this approach we end up with a number of images proportional to the length of the robot trajectory, and hence these view images do not satisfy our size criterion.

Related is the work by Ranganathan and Dellaert [15], where the goal is to identify a set of landmark locations, and then build a topological map using them. The images selected by this system, although small in number, and well suited to building topological maps, are however still not suitable for our purpose. First, this is an offline algorithm, which requires building a vocabulary of visual words [18] by clustering SIFT [12] features extracted from all the observed images. Second, we not only interested in selecting surprising landmark locations, but also images which represent the typical (i.e. mean) appearance of the world.

There is a body of literature on the related problem of offline video summarization, where we have random access to all the observed images. For example, Gong and Liu [7] video summaries were produced by exploiting a principal components representation of the color space of the video frames. They used a set of local color histograms and computed a singular value decomposition (SVD) of these local histograms to capture the primary statistical properties (with respect to a linear model) of how the color distribution varied. This allowed them the detect frames whose color content deviated substantially from the typical frame, as described by this model. Dudek and Lobos [4] used similar PCA technique, but also included coordinates of the images to produce navigation summaries. Ngo et. al. [14] first modeled the video as a complete undirected graph, and then used the normalized graph cut algorithm to partition the video into different clusters.

In our previous work [5], [6], we have looked at the problem of online navigation summaries, when the number of observations are known. In this paper we will focus on generating summaries when the number of observations is unknown or infinite.

We divide the problem of computing the summary into two different part. First, evaluating an image for its novelty and suitability for inclusion in the summary set, given the images already in the summary set. In Section II we first discuss this formally and then show how it can be implemented. Second, we need a sampling strategy which allows us to pick the best summary images. We discuss several such strategies in Section III

Y. Girdhar is a PhD Candidate in the School of Computer Science at McGill University, Montreal, Canada. yogesh@cim.mcgill.ca

G. Dudek is a Professor in the School of Computer Science at McGill University, Montreal, Canada. dudek@cim.mcgill.ca

II. SURPRISE

A. Bayesian Surprise

Itti and Baldi [9] formally define Bayesian surprise in terms of the difference between posterior and prior beliefs about the world. They showed that observations which lead to high Kullback-Leibler(KL) divergence [11] between posterior and prior visual appearance hypothesis, are very likely to attract human attention.

The relative entropy or KL divergence between two probability mass functions $p(x)$ and $q(x)$ is defined as:

$$d_{\mathrm{KL}}(p\|q) = \sum_{x\in\mathcal{X}} p(x)\log\frac{p(x)}{q(x)}. \tag{1}$$

KL divergence can be interpreted as the inefficiency in coding a random variable from distribution p, when assuming its distribution to be q.

In this paper we represent surprise with the symbol ξ.

$$\xi = d_{\mathrm{KL}}(\mathrm{posterior}\|\mathrm{prior}). \tag{2}$$

Suppose we have a set of summary images $\mathbf{S} = \{S_i\}$, which visually summarizes all our observations so far. Let F be a random variable representing presence of some visual feature. For example, F could represent presence of a given color, or a visual word [18], [17]. Let π^- be the prior probability distribution over all such features.

$$\pi^- = \mathbb{P}(F|\mathbf{S}) \tag{3}$$

Similarly, we can define the posterior probability distribution π^+, after observing a new image Z.

$$\pi^+ = \mathbb{P}(F|Z,\mathbf{S}) \tag{4}$$

Using Itti and Baldi's definition of surprise, we can then define *surprise* ξ in observing and image Z, given a summary $\mathbf{S}$ as:

$$\xi(Z|\mathbf{S}) \quad = \quad d_{\mathrm{KL}}\left(\pi^+\middle\|\pi^-\right). \tag{5}$$

Surprise $\xi(Z|\mathbf{S})$ can be interpreted as the amount of information gained in observing Z. Ideally, we would like to choose a summary set such that information gained after observing any random image from the terrain is small. In such a case, this would imply that our summary images already contain most of the information about the world.

B. Hypothesis Model

The above general definition of surprise is independent of the hypothesis model. For a realistic implementation, we must define how the images are described, and have a concrete description of the appearance model. Our work uses "visual words".

1) Drichlet over Visual Words: Sivic and Zisserman [18] have proposed a "bag-of-words model", in which each image is described as a histogram of word counts. The "words" used in the histogram are obtained by clustering SIFT [12] features. In related work, Ranganathan and Dellart [15], used an approximation of the Drichlet compound multinomial (DCM) [13] to build a measurement model, where each measurement is a bag-of-words histogram. Using this model, they compute Bayesian surprise, and identified landmark locations suitable to construct a topological map.

This approach to computing surprise is unsuitable for our purpose because of several reasons. First, it requires computing the visual "words" by clustering SIFT features extracted from all observed images. Hence it is only usable as an offline algorithm. Second, having a fixed predefined vocabulary implies fixed expressiveness, and hence fixed ability to detect surprises in an online setting.

2) Set Theoretic Surprise: Instead of modeling the appearance of the terrain with a single distribution over a static set of visual words, we propose to maintain a set of local hypotheses.

Hence, each image in the summary set has a corresponding distribution describing the probability of seeing a visual word or feature in that region. A set of these distributions can then be interpreted as the prior hypothesis. A definition of surprise using this set of local hypothesis can be computed in the following way [6]:

We define the prior hypothesis as a set of local hypothesis, each modeled by a distribution describing probability of seeing a visual features in the local region represented by a summary image.

$$\Pi^- = \left\{\mathbb{P}(F|S_1),\cdots,\mathbb{P}(F|S_k)\right\} \tag{6}$$

Similarly, we define the posterior hypothesis using the union of prior hypothesis set and the observation.

$$\Pi^+ = \left\{\mathbb{P}(F|S_1),\cdots,\mathbb{P}(F|S_k),\mathbb{P}(F|Z)\right\} \tag{7}$$

Now, analogous to Bayesian surprise, we would like to measure the distance between these two distribution sets. The Hausdorff metric provides a natural way to compute distance between two such sets. For two sets A, B, the Hausdorff distance between the sets is defined as

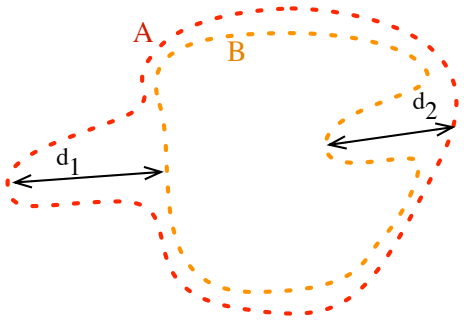
$$d_H(A,B) = \max\left\{\sup_{a\in A}\inf_{b\in B} d(a,b), \sup_{b\in B}\inf_{as\in A} d(a,b)\right\}. \tag{8}$$

Fig. 1 illustrates this graphically.

We define *Set Theoretic Surprise* ξ^* as the Hausdorff distance between the sets of posterior and prior distribution, with KL divergence as the distance metric.

$$\xi^*(Z|\mathbf{S}) \quad \stackrel{\mathrm{def}}{=} \quad d_{H,\mathrm{KL}}\left(\Pi^+\middle\|\Pi^-\right) \tag{9}$$

However since $\Pi^- \subseteq \Pi^+$, and only differs by one



$$d_1 = \sup_{a \in A} \inf_{b \in B} d(a,b)$$
$$d_2 = \sup_{b \in B} \inf_{a \in A} d(a,b)$$
$$d_H(A,B) = \max(d_1, d_2)$$

Fig. 1. *Hausdorff Metric* $d_H(A,B)$ measures the distance between two sets A and B.

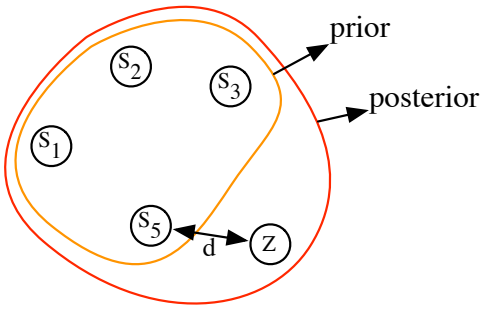


Fig. 2. *Set Theoretic Surprise*. We model our prior using a set containing the summary images $\{S_i\}$, and the posterior using a set containing summary images and the observed image Z. Each image is represented as a bag-of-words histogram, normalized to form a probability distribution. The surprise is then defined as the Hausdorff distance between these sets of probability distribution. We use KL divergence as the distance metric. Since the two sets only differ by one element, the Hausdorff distance can be simplified to only finding the closest element in the summary set.

element, expanding 9 we get:

$$\xi^*(Z|\mathbf{S}) = \max\left\{ \sup_{\pi^+ \in \Pi^+} \inf_{\pi^- \in \Pi^-} d_{\mathrm{KL}}(\pi^+ \| \pi^-), 0 \right\} \quad (10)$$
$$= \sup_{\pi^+ \in \Pi^+} \inf_{\pi^- \in \Pi^-} d_{\mathrm{KL}}(\pi^+ \| \pi^-) \quad (11)$$
$$= \inf_{\pi^- \in \Pi^-} d_{\mathrm{KL}}\left(\mathbb{P}(F|Z) \middle\| \pi^-\right) \quad (12)$$

This is visualized graphically in Fig.2.

3) Computing Vocabulary: Set Theoretic Surprise computation, as described above requires the computation of surprise in observing an observation image Z, given one of the summary image S_i. To do this we extract SURF [8] features from both the observation and the summary image, cluster them using the k means clustering algorithm, to generate a vocabulary of 64 words.

The normalized frequency count of these SURF words in the observation and the summary are then assumed to be their respective descriptions.

III. SAMPLING STRATEGIES

Given a surprise function, the task of picking k summary images can be performed in two different ways. If we know the total number of images that will be observed, then we can use the k Secretaries Hiring strategy proposed in [5], [6]. However, if we do not know the number of observations, or if the number of observations is infinite, then we must use a "Lake Wobegon Hiring strategy" [3] to continuously pick candidate summary images, and at the same time discard those which are not good.

A. Generalized k Secretaries Hiring Srategies

We would like our summary images to consist of k recent images which are most surprising. In our previous work [5], we considered the case where the length of the observation interval was constant. We showed that if we assume the score of each image to be from an unknown distribution, then the best online sampling strategy for a summary of size k is the following. Observe the first $n/(ke^{1/k})$ images, where n is the total number of observations; and then set the threshold to the maximum score in this observation interval. Now, simply choose the first k images which exceed this threshold. This threshold optimizes the probability of selecting all k top scoring images, with probability of success approaching $1/(ek)$ [5]. This strategy works in the case where the score of an image is independent of the summary set. In our case, however, once we add an image to the summary set, the surprise function used to compute the novelty is altered, making the performance results inapplicable.

In [6], we proposed a modification of the fixed threshold strategy above, where we recompute the threshold after the selection of each summary image. After each selection, we recursively run the selection algorithm on the remaining observation, for selecting the remaining summary images. This strategy, however, again requires us to know the size of the observation set.

B. Lake Wobegon Hiring Strategies

Broder et. al. [3] named the strategy of picking samples above the mean or median score as "Lake Wobegon" strategies [1].

If the number of observed images is unknown or possibly infinite, then picking images above the mean surprise of previously selected images is a trivial strategy. Moreover, we only consider selecting images which have locally maximum surprise. However, if we continuously select the images which are above the average surprise of the previously selected images, and assuming the surprise scores are from a uniform random distribution, then it can be shown that we will have infinitely many images, as time $\to \infty$ [3].

Since we are interested in maintaining a finite number of images in the summary set, we then must come up with a strategy to discard a summary image, when the summary size exceeds the max size. We propose a few different discarding strategies, each of which leads to a different kind of a summary:

[1]Named after the fictional town "Lake Wobegon", where according to the Wikipedia "all the women are strong, the men are good looking, and all the children are above average."[3]

1) *Discard Oldest:* We define the age of a summary image in terms of the time of the last observation which matched that summary image. Hence, if a summary image is regularly observed, it is kept in the summary. This ensures that we have images which correspond to the mean appearance of the world in our summary, since they are needed to identify the surprises. This strategy produces a summary which focuses on describing the recent observations.
2) *Discard Least Surprising:* Discard the summary image which is least surprising, given the remaining summary images. If S_r is the discarded summary image, then we have:

$$r = \underset{i}{\operatorname{argmin}}\, \xi^*(S_i | \mathbf{S} - \{S_i\}). \tag{13}$$

Discarding least surprising image is a good strategy if we would like a temporally global summary of all the observations seen so far.
3) *Hybrid:* One could also easily think of a many ways to combining the above two strategies. For example, one way is to consider the average rank ordering of an image among the above two strategies, and then discard the image with the highest average rank ordering.

IV. IMPLEMENTATION

We implemented the proposed summarization framework using C++, and OpenCV [2]. Our current implementation allows us to process video frames of size 640x480 in real time, at about 1 frames per second on a Core 2 processor.

We tested our approach from video footage collected by two different robots. We used the AQUA [16] amphibious robot to collect video while swimming swimming over a coral reef. We also used a UAV made by Procerus Technologies, equipped with a GPS and Autopilot to collect footage from an altitude of about 100 meters.

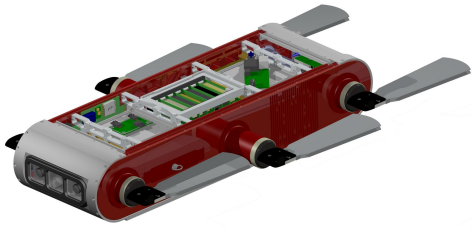
V. RESULTS

We present results of generating summaries from two different type of environments: terrain as seen by an aerial vehicle and underwater over a coral reef.

A. Terrain Summaries

We flew our aerial vehicle over the region shown in Fig. 4(b). The downward looking video from the plane was then fed into our summary program. Fig. 4(c) shows the evolution of the summary set over time. We start off with the putting the first observed image in the summary set, as indicated by the first row of Fig. 4(c). Now for each new observation, we compute its surprise score given the images already in the summary set. If the surprise is above the threshold, which is initially set to zero, we then include the image in the summary set. Each successive row of Fig. 4(c) shows the state of the summary set after 3 modifications. The final row is the summary after observing the last image.

We used the "discard least surprising" strategy to eliminate an image from the summary set when it grows larger than the maximum specified size of 6. Due to lack of space, we do not present results from other strategies.



(a) AQUA Amphibious Robot

(b) Procerus UAV

Fig. 3. Robots and used for generating visual summaries.

Fig. 4(a) show the surprise of the incoming observations and acceptance threshold overtime. We see that initially, since the threshold is low, we rapidly pick several images and in the process the threshold also grows rapidly. This is also clear from Fig. 4(c), where we see the initial rows are filled with similar looking images, which is result of a low threshold.

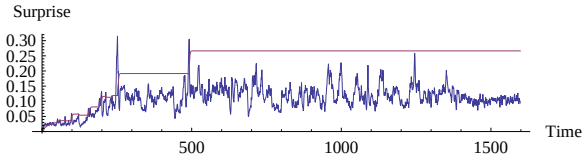
B. Underwater Summaries

The evolution of the summary set using the underwater video data is shown in Fig. 5(b). About 1200 observations were processed. The final summary shown in the last row of the figure. The summary was generated using the same parameters as the aerial summary. The main difference in behavior is that we see the summary images to be quite different right from the beginning. This is also apparent from continuous regular stepping of the threshold as seen in Fig 5(a). The behavior is explained by the continuous variation in terrain appearance in this data set.

VI. CONCLUSION

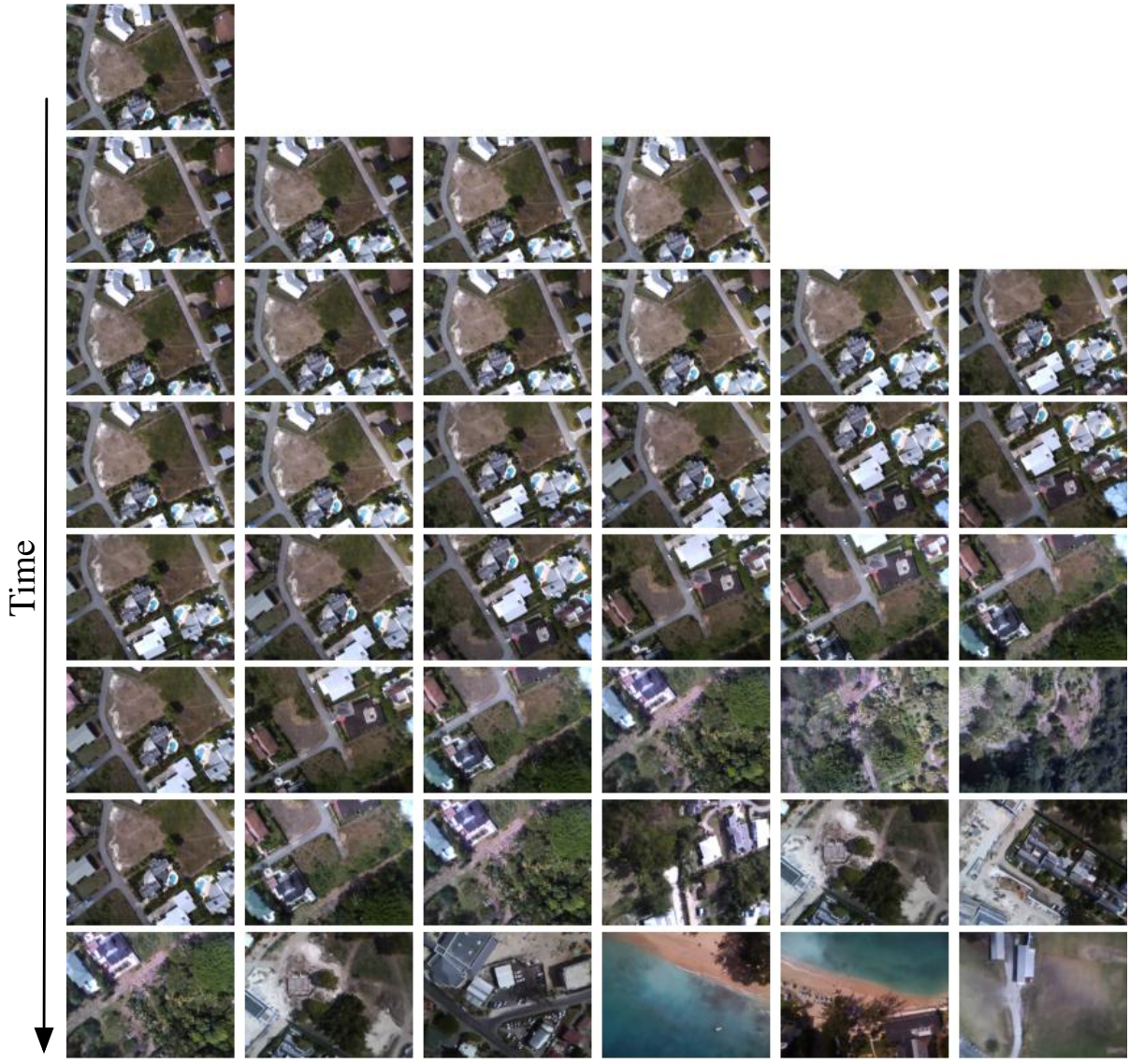
In this paper we have described an on-line a technique for generating video summaries in real time. To the best of our knowledge, this is the first technique which deals with the summarization problem in the case where the process operates incrementally and online, including the progressive evolution of a domain dependent image representation, and which operates on video sequences on unbounded length.

The general strategy followed by our algorithm is to compute the "surprise" of incoming observation and include it in the summary set if it is above a threshold value. Itti and Baldi formally define Bayesian surprise as the KL divergence between the posterior and prior hypothesis. We model the



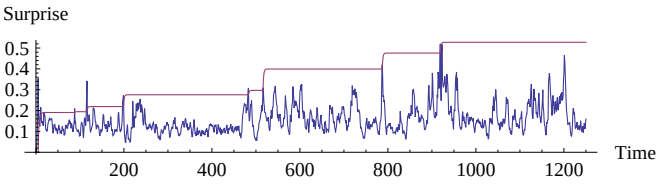
(a) Surprise and selection threshold over time.

(b) Aerial photo of the entire region.

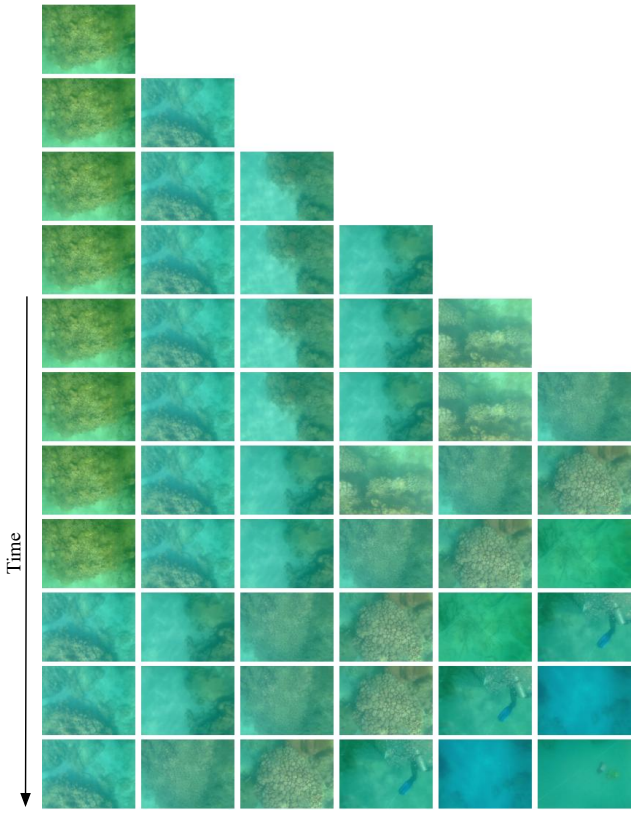


(c) Summary Evolution

Fig. 4. Succession of 6-frame navigation summaries computed by the system at successive points in time during the flight of an aerial robot. Each row depicts an intermediate navigation summary computed based on the (partial) data recorded as the robot captures additional frames. Time progresses downwards with the top being the first result and the bottom row being the last. The first 4 or 5 rows are based on video captured over a small region and roughly uniform terrain. Subsequent frames describe increasingly varied types of images including frames that have land covering the North half of the image, or land covering only the South half of the image. (See text for further details.)



(a) Surprise and selection threshold over time.



(b) Summary Evolution

Fig. 5. Summary generated by an underwater robot as it traverses a coral reef. (a) show the surprise of a new observation given the images in the summary set at that time. (b) show the evolution of the summary set over time. Each row corresponds to an instance of the summary set. The final row is the final summary.

prior hypothesis using the images in the summary set. The posterior is modeled as the union of the images in the summary set and the incoming observation.

Each image is described using the bag-of-words technique. The vocabulary used is computed by clustering SURF words from the summary and the observation. We do not use a global set of static words and hence do not require any prior training.

While are results so far have been satisfying and useful, several open problems remain to be examined. A recurring issue in problems like this is the choice of suitable representations, the relationship to human performance, and the connection to semantic information of task specific constraints. For some summarization tasks, specific aspects of a domain may be particularly important, and we are examining how these can be specified and incorporated into our process.

Moreover, due to the online nature of the presented algorithm, it has applications beyond just video summarization which we hope to explore in the future. The event of including an observation in the summary set can be used to trigger actions such are directing robot motion, or dropping sensors.

REFERENCES

[1] J. Bauer, N. Sunderhauf, and P. Protzel. Comparing several implementations of two recently published feature detectors. In *Proceedings of the International Conference on Intelligent and Autonomous Systems, IAV, Toulouse, France*, 2007.

[2] G. Bradski. The opencv library. *Doctor Dobbs Journal*, 25(11):120–126, 2000.

[3] Andrei Z. Broder, Adam Kirsch, Ravi Kumar, Michael Mitzenmacher, Eli Upfal, and Sergei Vassilvitskii. The hiring problem and lake wobegon strategies. In *SODA '08: Proceedings of the nineteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1184–1193, Philadelphia, PA, USA, 2008. Society for Industrial and Applied Mathematics.

[4] Gregory Dudek and John-Paul Lobos. Towards navigation summaries: Automated production of a synopsis of a robot trajectories. In *Canadian Conference on Computer and Robotic Vision(CRV)*, Kelowna, British Columbia, May 2009.

[5] Yogesh Girdhar and Gregory Dudek. Optimal online data sampling or how to hire the best secretaries. In *Canadian Conference on Computer and Robot Vision*, Kelowna, British Columbia, May 2009.

[6] Yogesh Girdhar and Gregory Dudek. Online navigation summaries. In *IEEE International Conference on Robotics and Automation (ICRA)*, May 2010.

[7] Yihong Gong and Xin Liu. Video summarization using singular value decomposition. In *Proc. of CVPR*, pages 174–180, 2000.

[8] Luc Van Gool Herbert Bay, Tinne Tuytelaars. Surf: Speeded up robust features. *ECCV*, 2006.

[9] Laurent Itti and Pierre Baldi. Bayesian surprise attracts human attention. *Vision Research*, 49(10):1295 – 1306, 2009. Visual Attention: Psychophysics, electrophysiology and neuroimaging.

[10] K. Konolige, J. Bowman, JD Chen, P. Mihelich, M. Calonder, V. Lepetit, and P. Fua. View-based maps. *RSS'09*, 2009.

[11] S. Kullback. *Information theory and statistics*. John Wiley and Sons, NY, 1959.

[12] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2), 2004.

[13] R.E. Madsen, D. Kauchak, and C. Elkan. Modeling word burstiness using the dirichlet distribution. In *Proceedings of the 22nd international conference on Machine learning*, page 552. ACM, 2005.

[14] Chong-Wah Ngo, Yu-Fei Ma, and Hong-Jiang Zhang. Video summarization and scene detection by graph modeling. *IEEE Transactions on Circuits and Systems for Video Technology*, 15(2):296–305, 2005.

[15] A. Ranganathan and F. Dellaert. Bayesian surprise and landmark detection. In *Proceedings of the 2009 IEEE international conference on Robotics and Automation*, pages 1240–1246. Institute of Electrical and Electronics Engineers Inc., The, 2009.

[16] Junaed Sattar, Gregory Dudek, Olivia Chiu, Ioannis Rekleitis, Philippe Giguère, Alec Mills, Nicolas Plamondon, Chris Prahacs, Yogesh Girdhar, Meyer Nahon, and John-Paul Lobos. Enabling autonomous capabilities in underwater robotics. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS*, Nice, France, September 2008.

[17] Josef Sivic and Andrew Zisserman. Video google: A text retrieval approach to object matching in videos. In *In Proc. ICCV*, pages 1470–1477, 2003.

[18] Josef Sivic and Andrew Zisserman. Video google: Efficient visual search of videos. In *In Toward Category-Level Object Recognition, volume 4170 of LNCS*, pages 127–144. Springer, 2006.