VSRV: Video Summarization for Rushes Videos

Tiago Oliveira Cunha, Flávio Gonçalves Henriques de Souza, Gisele Lobo Pappa, Arnaldo de Albuquerque Araújo
Departamento de Ciência da Computação - DCC
Universidade Federal de Minas Gerais (UFMG)
Belo Horizonte, Brasil
Email:{tocunha,sflavio, glpappa,arnaldo}@dcc.ufmg.br

***Abstract*—Producing, storing and distributing video content on the Web was never so quick and easy. For this reason the availability of multimedia data is increasing very fast and generating a great demand for new methods to explore the content available in these data. This has been the goal of automatic video summarization, one the most studied task in video processing and understanding. A video summary provides a short version of the original video without losing the central idea. Here we focus in one method for automatic rushes video summarization. Rushes consist of unedited material generated during the recording of a video film, and have a special structure characterized by a high number of repetitions and a great number of useless segments. To solve this problem, we propose an approach based on spatial and spatial-temporal features represented by a bags-of-visual-words. This representation is robust to a series of transformations in image and occlusion. The task is modeled as an optimization problem, and a multiview learning strategy is applied. Results on the BBC Rushes database were compared with the three best methods submitted to the TRECVID, and showed the methodology to be promising for dynamic rushes video summarization.**

***Keywords*-Video Summarization;rushes video;**

I. Introduction

With recent advances in technology, production, storage and distribution of video content has never been so quick and easy. But is not only ordinary home-users that are producing a record number of multimedia videos. The film making industry is also producing as much material as ever.

Making efficient use of video information requires that data to be accessed in a user-friendly way. For this, it is important to provide users with a concise video representation to give an idea of a video content, without having to watch it entirely, so that a user can decide whether watch the entire video or not, saving time and effort. This has been the goal of a quickly evolving research area known as video summarization [1].

A video summary can be of two types: static (a sequence of key-frames) or dynamic (a sequence of video segments). The dynamic version has the advantage of be able to incorporate audio and movement aspects, which might be more attractive to the users.

This paper focuses on dynamic summarization of rushes videos [2]. Rushes are the raw material used to produce a video, and their summarization has particular characteristics, for example, several redundant sequences and sequences used only for marking and separating recordings, named junk shots.

Many different approaches can be found to generate summaries of rushes videos [3], [4], [5], [6]. The great majority is based on clustering techniques used for redundancy removal. However, as these approaches are based on distances among video segments, the segments need to be effectively characterized, in a way that distance metrics can really identify similar segments. Several strategies can be used to characterize a segment, but there is no consensus on what are the best features to be employed.

Here, we present VSRS[1] an approach for dynamic rushes videos summarization based on spatial and spatial-temporal features, represented by a bags-of-visual-words approach (BoW).

Our system selects important segments of video, trying to identify non-redundant segments containing high motion activity. The importance of a segment is measured according to the assumptions in [5], which says that more movement represents more information. The system also eliminates uninteresting segments, including colorbars and clapperboards.

The BoW approach is an un-structured global representation of videos which is built using a large set of local features. In BoW, descriptors extracted at numerous locations in space and time are clustered into a number of visual words and the video is represented by a histogram of these words.

BoW tries to reduce the semantic gap between low-level features and image visual content, and has been used in the literature in various scenarios of pattern detection and classification, achieving good results due its robustness to a series of transformations in the image and occlusion.

Besides representing segments using mid-level features generated from low-level features, the proposed methodology also employs a strategy inspired by multiview learning [7]. Multiview learning uses different representations extracted from a unique object (segment) and learns from each of them independently. Here we work with three views, represented by the following descriptors: SIFT, Hue-SIFT and STIP. Each descriptor is used to generate a BoW, and the learning process corresponds to finding the most similar segments (represented by BoW generated from the descriptors) using a clustering algorithm for each view separately.

Having clustered the segments, for each cluster, the most representative segment is extracted, and a summary is generated by modeling the summarization task as an optimization problem. This is necessary because the summaries have a maximum predefined duration. However, this time constraint

[1]Master's Thesis

has to be obeyed while preserving the most important video segments.

Hence, the final summary generation is modeled as a knapsack-problem, with the duration of the summary being equivalent to the weight of the knapsack and the amount of movement on each segment the benefit of an item. Finally, after summaries for each view are created, they are united to form a final summary, using the same approach just described.

The summaries generated are evaluated by a set of users following the evaluation methodology proposed in the TRECVID BBC Rushes Summarization Task [2]. Results were compared with the three best approaches submitted to TRECVID, and showed that the method is competitive with all of them considering three out of four metrics evaluated.

The remainder of this paper is organized as follow. Section 2 describes related works in the area of dynamic rushes videos summarization. Section 3 introduces our methodology, while Section 4 reports experimental results. Finally, Section 5 draws some conclusions and discusses future works.

II. Related Work

Automatic video summarization is facing a fast development in recent years, and is becoming a comum tool in most of the multimedia management systems to help users to save time in video database analysis. In general, these techniques: (i) employ low-level video features, (ii) pay special attention to the length of the summary and (iii) make use of all information available for the shot. Additionally, some works perform an extra smoothing step to make the final summary look more natural. In the specific case of rushes videos, most of the approaches also use clustering techniques to remove redundancy, employ a importance measure to choose sequences to form the summary, and remove junk shots.

This section describes the three most effective methods selected from the 22 proposed to solve the TRECVID 2007 task. These methods will be later used as baselines for comparisons for the methods proposed here.

Detyniecki and Marsala [4] (*Lip6*) developed a summarization approach called *stacking*. It is based on shot boundary detection and elimination of redundant content. In this method, shots are compared and, from the most similar ones, only the longest is selected. Following that, a technique for adaptive acceleration of the video frames is applied. A informative measure based on visual similarity among the frames of selected shots is defined, and the most important shots displayed at the standard rate of frames per second, while the shots considered non-informative are exposed to a higher rate of frames per second. One problem with this method is that it does not perform junk shots detection. Although very simple, this approach was also selected as one of the three best in TRECVID 2007. In particular, it offers easy to understand summaries that keep most of the original information, meets the target compression rate, and has average scores of redundancy perception.

Another video summarization approach was proposed in [8] (*CityU*), the summarization approach is based on a complex and detailed videos analysis. This analysis includes shot boundary detection, object detection, camera movement estimation, matching and tracking of interesting regions, audio classification and speech recognition. A representative measure was used to model object presence and four audiovisual events (object movement, camera movement, scene changing and speech content) in video segments. The segments with greater representativeness were chosen to compose the summary. The detection of junk shots was done by comparing the color features with predefined junk patterns. In contrast with other traditional methods in the literature, this one performs a lot of processing on the video structure.

Finally, [3](*Nii*) proposed an approach based on local color characteristics and grouping for rushes video summarization. Initially, the video is divided into segments that present high visual similarity. The central frame of each segment is extracted as a key-frame. The key-frames are then grouped, and from each group the segment that has the longer duration is selected. Finally, based on the length of the summary, samples of the segments are selected. As in [4], this work does not perform junk shots detection.

III. VSRV Approach

This section presents the VSRS, an approach to summarize rushes videos. The approach deals with the very specific structure of rushes videos and use spatio-temporal features represented by a BoW approach to produce the summaries.

Figure 1 shows the general scheme of our methodology. Initially, the video is segmented into basic units (segments), and the segments that do not have relevant information to the summary are eliminated. The remaining segments are described using three descriptors, one spatial-temporal (STIP [9]) and two spatial (SIFT [10] and Hue-SIFT [11]). Each description is then represented using a bags-of-visual-words approach (BoW) [12] (Section III-A).

Having three processing sequences (one for each BoW description), the K-means algorithm is run to find the most similar segments. For each cluster created, only the most representative segment is extracted, once repeated segments are usually different takes of a single scene (Section III-B). Next, a summary is generated by modeling the summarization task as an optimization problem, where the summary duration (predefined) and the most important video segments are optimized (Section III-C). Finally, after summaries for each descriptor are created, they are united to form a single and final summary, using the same approach described above.

Here, the method used to segment the videos is an adaptation of the approach proposed in [5], and was chosen due to its low computational cost and good results achieved in terms of shot boundary detection and junk shots elimination [2]. It benefits from the use of local color histograms and motion vectors, and is done in three phases: (i) shot boundary detection, (ii) subshot detection, and (iii) junk shots detection. Recall that in rushes one shot represents a scene recording, and we might have many shots of the same scene.

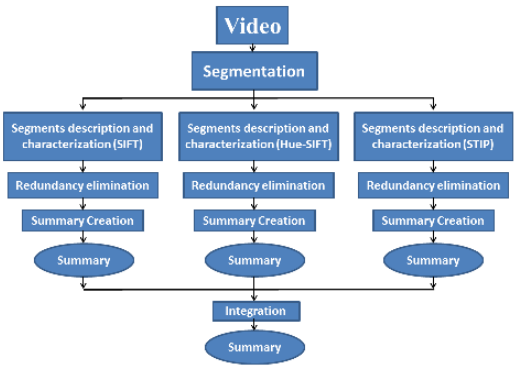

Fig. 1. Architecture of the proposed approach.

A. Segments Description and Characterization

In the proposed method, the basic video units are described by two spatial descriptors and one spatial-temporal descriptor. These descriptors detect and describe interesting points in an image or video. An interesting point is a specific point (or region) that presents significant intensity variation in more than one direction. They are widely used in computer vision tasks like tracking and recognition [13].

In order to represent the segments previously defined we use a bags-of-visual-words approach (or bags of visual features, or bags of keypoints) (BoW). BoW [12] is a model free technique, where no previous knowledge is given about the domain, and is inspired by traditional techniques of textual information retrieval. BoW is a robust representation for images where each image is seen as a set of regions or points where only the visual information of the region matters, and no information about the location of the point in the image is needed. These points are called visual words.

The BoW method is executed in four steps. First, a method to detect and describe the points of the image has to be applied. The descriptors extracted from the image need to be invariant to changes that are irrelevant to the categorization task (image transformations, lighting variations and occlusions) but rich enough to discriminate categories.

The video segments are described using the STIP [9], SIFT [10] and Hue-SIFT [11] descriptors, which have the characteristics just mentioned. SIFT and Hue-SIFT are spatial descriptors that work with still images, so we adopt a strategy of 2D representation for the subshots. We took advantage of the fact that all the frames in a subshot have great visual similarity, and extracted the central frame as a key-frame to represent the subshot as done in [3]. Hence, the subshots are described by the STIP and the key-frames by the SIFT and the Hue-SIFT.

In a second step, a vocabulary is defined. This definition is based on the choice of a set of interesting points (visual words), which is done randomly from all available points. Next, the third step is to associate each descriptor to a visual word in the vocabulary. This association is done by calculating the Euclidian distance between the points of the image and the vocabulary. The closest visual word in the vocabulary to an image point is stored in order to generate a histogram of occurrence of visual words.

Bow provides a image representation more informative in terms of low-level features, as it is expected that the characteristics that compose the histograms are indeed representative patterns to describe the image content.

B. Redundance Elimination

In order to eliminate redundancy, the well-known k-means algorithm is applied to generate clusters. The idea is that, after the clustering, BoWs representing similar segments will belong to the same group, and from each group only one representative segment is chosen. The number of groups K is chosen according to the number of video shots. We use this strategy to get an estimation of the number of video scenes, once in rushes one shot normally corresponds to a shooting of a scene.

The choice of the most representative segment per group is done by calculating the proximity of the BoWs to the center of the group, and the BoW closest to the center is chosen to represent the group. The proximity is calculated using the Euclidian distance.

C. Summary Creation

Finally, we have a list of segments and need to keep just the most informative ones. The summary duration is defined apriori. To ensure that all summaries are in accordance with a predefined maximum duration, the problem is modeled as the well-known binary knapsack problem [14]. Given a set of *n* objects and one knapsack, where: c_j is the benefit of the object j, w_j is the weight of the object j, and b is the capacity of the knapsack. The problem is to determine which objects should be placed in the knapsack to maximize the benefit in such a way that the weight of the knapsack does not exceed its capacity. Our goal is formally defined as: Maximize $z = \sum_{j=1}^{n} c_j s_j$, subject to $\sum_{j=1}^{n} w_j s_j \leq b$ with $s_j \in \{0, 1\}$.

We want to generate summaries formed by segments with higher motion values and with a duration that does not exceed a predetermined value. This is because we assume that the larger the motion value of a segment is, the more information it provides [5]. Hence, we define that the number of frames corresponds to the weight of the knapsack and the amount of movement correspond to the benefit. The resolution of the knapsack problem give us the set of segments with higher motion and within the limited summary time. The knapsack problem is solved using the dynamic programming method [14].

1) Integration of the Summaries: One of the contributions of this work is the use of three descriptors to characterize the video segments. It was inspired by the idea of multiview learning [7], which is a machine learning setting that explicitly exploits a set of disjoint features, each one sufficient to learn the target concept. The idea is that the features are complementary and generate better results than those obtained with a single description.

In multiview learning each description is named vision. Ideally the visions should be uncorrelated and disjoint, but in real world databases this is very hard to reach. For each vision a classifier is learned, and at the end of the process of classification, the results generated by all the classifiers are combined in order to reduce the classification error [7].

Having three summaries generated from three different descriptors, and want to generate a fourth summary from them, which is expected to have higher quality than the three separate ones. We first unite the three summaries, and if there is an overlap between two segments, a new segment is formed using the lower and upper bounds of the segments, and a new motion value is calculated. Finally, the same steps used before to generate independent summaries are reapplied: redundancy elimination and selection of the most informative segments.

IV. Experimental Results

Creating a video summary is a hard task, and finding methodologies for evaluating their quality is even harder. Studies have showed that even when a human summarizes the same text twice, he/she will usually not agree with him/herself. Comparing two video summaries created by different methods has this same problem. Although the problem of summarization is being intensively investigated, there is no ideal method for assessing the quality of summaries. Some efforts are being made to create a standard evaluation approach, as proposed in the video summarization task of TRECVID 2007 [2].

Aiming to make evaluation easier, the database used in the experiments is the same used in TRECVID 2007 [2], which provided 42 videos associated with a ground truth. The ground truth consists of textual descriptions of important scenes that should be included in a good summary. In the annotations there is always a concern to specify camera angle, distance or other information which makes the video segment unique.

Four measures of summary quality are evaluated: (i) inclusion of ground truth content, (ii) junk shots presence, (iii) redundant segments presence and (iv) time of judgment. The summaries are evaluated in a subjective way by users. The evaluating methodology is based on the proposed in the TRECVID [2], where human assessors watch the summaries and give their opinion with respect to the amount of relevant information contained in the summary. Together with the summaries a list of topics that represent important video segments is provided to each assessor.

Each summary was judged by three users that evaluate the presence of 12 topics randomly chosen from the list extracted from the ground truth. The process of finding the list of topics in the summary is timed to compute a measure of effort in the judgement. The evaluation process also collects measures of usability/satisfaction. To assess the amount of redundant segments and junk shots present in the summary a scale ranging from 1 to 5 was defined, where 1 is strongly agree and 5 strongly disagree. A complete description of the evaluation methodology can be found in [2].

These metrics were used to evaluate a total of seven methods, including the four types of summaries produced in this work, from now on referred as R-SHS (produced by integration of SIFT, Hue-SIFT and STIP), R-SIFT (produced by SIFT), R-Hue (produced by Hue-SIFT), R-STIP (produced by STIP). The other three methodologies, used as baselines, are described in Section 2: CityU (produced by[8]), Lip6 (produced by [4]) and Nii (produced by [3]). These three works are the best evaluated in TRECVID 2007 in terms of inclusion of ground truth content, where they did not show statistical difference.

TABLE I
Average Results for the four metrics assessed: ground truth inclusion, junk shot detection, redundant segments and evaluation time

	Inclusion(%)	Junk	Redundant	Time
R-SHS	57,11	4,28	2,72	126,02
R-SIFT	52,76	4,23	2,84	81,08
R-HueSIFT	54,44	4,29	2,83	92,67
R-STIP	53,83	4,31	2,88	79,61
CityU	53,97	4,50	3,70	109,79
Lip6	49,07	2,60	2,80	82,03
Nii	58,63	2,27	2,75	130,99

In the next section, the results for the four evaluated metrics are presented in terms of boxplots and means, and an hypothesis test is made to compare the values obtained by the methodologies in accordance with each measure evaluated. We use the t-test [15], with confidence of 1 - α (1 - α = 0.95).

A. Inclusion of Ground Truth Content

Starting with the inclusion of ground truth content, column *Inclusion* in Table I shows the average percentage values obtained by the approaches according to the mean of three evaluators assessing 42 videos. The values range from 49,06 (Lip6) to 58,63 (Nii), showing the great subjectivity of the summarization problem. Although Nii makes just a low-level video analysis, it gets the best average results for the inclusion of ground truth content. Initially, we expected CityU to obtain the best results due to its detailed video analysis. The worst average results are achieved by Lip6. The reason for that might be related to the fact that it chooses the segments to compose the summary by visual similarity, and apply an adaptive acceleration. Hence, it can make some scenes hard to understand. The results also show that the summaries produced by the multiview learning inspired strategy obtained better average results that those observed in the summaries produced by the descriptors separately.

Figure 2 shows the boxplots with values achieved by each system methodology for inclusion of ground truth content. The boxes on the graph represent the intervals between the first and third quartiles of the samples, where 50 % of the measurements can be found. The bold line in the box indicates the median value of the data. If the median line within the box is not equidistant from the extremes, it says that the data are asymmetrical. In the extremes of the chart are dotted vertical lines, which indicate the minimum and maximum values, unless outliers are present. In the latter case, graphs extends to a maximum of 1.5 times the interquartile distance.

The points out of the graph are the (suspected) outliers. Finally, the dotted horizontal lines represent the confidence interval.

Since there is an overlap of the confidence interval of all methodologies, the analysis of Figure 2 does not provide evidence to conclude if there is a statistical difference among the methods. Hence, we performed a hypothesis test for a more detailed analysis of the differences among the measures.

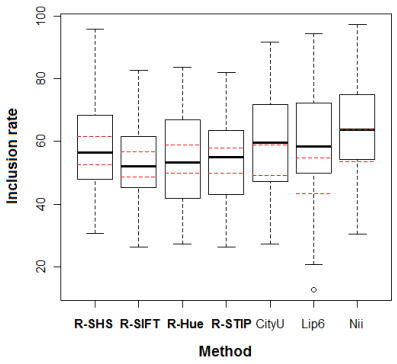

Fig. 2. Percentage of inclusion of ground truth

According to the hypothesis test, in only two cases (R-SHS with R-SIFT and R-SHS with R-STIP) we observe values smaller than the significance level α, and in these two cases the null hypothesis is rejected. Thus, we can conclude with 95% of confidence that the R-SHS is statistically superior when compared to the R-SIFT and R-STIP methodologies regarding the rate of inclusion of ground truth content. For the other methodologies we do not have evidence to conclude if there are statistical differences among them.

B. Presence of Junk Shots

For evaluating the presence of redundant segments, the same scale ranging from 1 to 5 introduced before was used, where 1 is strongly agree and 5 strongly disagree. Column *Junk* in Table I shows that the summaries produced in this work and the summary generated by CityU have better average results than the summaries generated by Lip6 and Nii. This result was already expected, since Nii and Lip6 do not perform any junk shots detection.

Figure 3 shows the boxplots for the presence of junk shots. The fact that there is no overlapping among the confidence intervals of the Nii and Lip6 with other confidence intervals gives a clue that there are statistical differences when comparing these two methodologies with the others. The hypothesis test showed with statistical confidence that values obtained by Lip6 and Nii are worse than those produced by other methods for the presence of junk shots, and we cannot say if there is statistical difference among our method and CityU.

C. Presence of Redundant Segments

Column *Redundant* in Table I presents the average values to each methodology regarding the presence of redundant segments, and Figure 4 shows the boxplots for this measure. We can see that the summaries produce by CityU obtained the

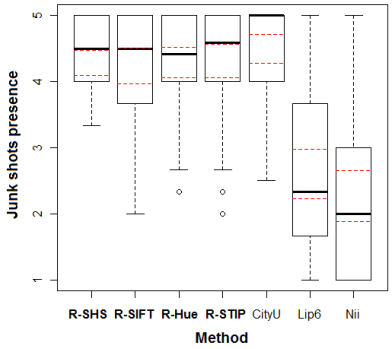

Fig. 3. Junk Shots Presence

best average results, while the other methods obtained similar results among them. In the TRECVID 2007 evaluation this measure had a high correlation with the rate of inclusion of ground truth. This happened because the presence of identical segments made it easier for the user to find the ground truth in the video, increasing the rates of inclusion. This trend was also observed here, as the methodology that achieved the best average results for the inclusion of ground truth content also obtained the worst results for the absence of redundant segments.

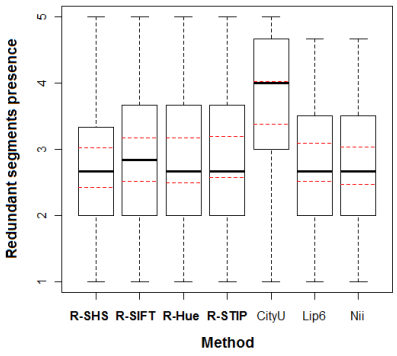

Fig. 4. Redundant segments presence

The hypothesis test confirmed with 95% of confidence that CityU is superior to the proposed methodology, Lip6 and Nii regarding the presence of redundant segments. This is explained by the fact that CityU uses a complex technique based on graph search and several color characteristics for the removal of redundancy, while other methodologies use only visual similarity based on local color histograms. For the comparison among the other methodologies, the null hypothesis cannot be discarded, and hence we do not have evidence to ensure statistical differences among them.

D. Evaluation Time

Among the metrics evaluated this is the most subjective because it is totally dependent of the user. It was chosen to be part of the assessment to give a sense of how ease it is for the user understand the summary. This metric showed the greatest amount of outliers and data dispersion. Identifying

a clear reason for this behavior is not simple, as users where not in a completely isolated environment and could be distracted during evaluation. On one hand, we expected that the methods using acceleration need a greater amount of time for evaluation, but this was not observed. On the other hand, we identified that the methodologies that obtained the best average rates for inclusion of ground truth were also those with the highest evaluation time.

The column *Time* in Table I shows the mean values of the time of evaluation for each methodology. The values ranged from 79,61 (R-STIP) to 126,02 (R-SHS) seconds. Looking at Figure 5, Nii and R-SHS obtained the greatest time of evaluation, while other methodologies had similar measures.

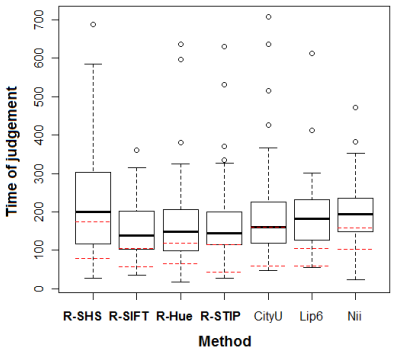

Fig. 5. Time of judgment by users (in seconds)

V. Conclusion

This work introduced a VSRS, a video summarization approach for rushes videos. The method selects important segments of video, trying to identify non-redundant segments containing high motion activity, also eliminates uninteresting segments, including colorbars and clapperboards.

The method is based on spatial-temporal features represented by a bags-of-visual-words approach (BoW). BoW tries to reduce the semantic gap between low-level features and image visual content, and was implemented over the SIFT, Hue-SIFT and STIP descriptors.

The method also employed a strategy inspired by multiview learning, where independent summaries for the three descriptors were generated separately, and later merged to provide a unique and final summary. To ensure that the summaries were formed by the most important video segments and had a predefined duration, the summary generation task was modeled as the knapsack problem, where we used the assumption that segments with higher motion provide more information.

Experiments with TRECVID BBC rushes database showed that our methodology is competitive with CityU, Lip6 and Nii, which were the three best evaluated methodologies in TRECVID 2007 summarization campaign in terms of inclusion of ground truth content. The methods were evaluated using four measures based on the evaluating methodology proposed in the TRECVID competition. For three of these measures, there is no statistical evidence to show that the summaries produced by this methodology are better than those produced by CityU, Lip6 and Nii. The results also showed that the summaries produced by multiview learning based approach were statistically superior to the ones produced by the descriptors separately in relation to the measure of rate of inclusion of ground truth.

During the analysis of the results, we identified a set of improvements which can lead our methodology to better results. They involve, for example, the use of others types of information to describe the video segments (text, audio and semantics). Applying other importance measures to choose the most representative segments. Finally, the method can be easily adapted to work with others video genres.

Acknowledgment

This work was partially supported by Inweb, CNPq, CAPES and FAPEMIG.

References

[1] J. Almeida, N. J. Leite, and R. da Silva Torres, "Vison: Video summarization for online applications," *Pattern Recognition Letters*, vol. 33, no. 4, pp. 397–409, 2012.

[2] P. Over, A. F. Smeaton, and P. Kelly, "The trecvid 2007 bbc rushes summarization evaluation pilot," in *Proc. of the Int. Workshop on TRECVID video summarization*, 2007, pp. 1–15. [Online]. Available: http://doi.acm.org/10.1145/1290031.1290032

[3] D.-D. Le and S. Satoh, "National institute of informatics, japan at trecvid 2007: Bbc rushes summarization," in *Proc. of the Int. Workshop on TRECVID video summarization*, 2007, pp. 70–73. [Online]. Available: http://doi.acm.org/10.1145/1290031.1290044

[4] M. Detyniecki and C. Marsala, "Video rushes summarization by adaptive acceleration and stacking of shots," in *Proc. of the Int. Workshop on TRECVID video summarization*. ACM, 2007.

[5] C.-M. Pan, Y.-Y. Chuang, and W. H. Hsu, "Ntu trecvid-2007 fast rushes summarization system," in *Proc. of the Int. Workshop on TRECVID video summarization*, 2007, pp. 74–78. [Online]. Available: http://doi.acm.org/10.1145/1290031.1290045

[6] E. Dumont and B. Mérialdo, "Rushes video summarization and evaluation," *Multimedia Tools Appl.*, vol. 48, pp. 51–68, May 2010. [Online]. Available: http://dx.doi.org/10.1007/s11042-009-0374-9

[7] I. Muslea, S. Minton, and C. A. Knoblock, "Active + semi-supervised learning = robust multi-view learning," in *Proc. of the 19th Int. Conf. on Machine Learning*, 2002, pp. 435–442. [Online]. Available: http://portal.acm.org/citation.cfm?id=645531.655845

[8] "THU-ICRC at rush summarization of TRECVID 2007," in *Proc. of the Int. Workshop on TRECVID video summarization*. ACM, 2007.

[9] I. Laptev, "On space-time interest points," *Int. J. Comput. Vision*, vol. 64, pp. 107–123, September 2005. [Online]. Available: http://portal.acm.org/citation.cfm?id=1085595.1085605

[10] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision*, vol. 60, pp. 91–110, November 2004. [Online]. Available: http://portal.acm.org/citation.cfm?id=993451.996342

[11] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek, "Evaluating color descriptors for object and scene recognition," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1582–1596, 2010. [Online]. Available: http://www.science.uva.nl/research/publications/2010/vandeSandeTPAMI2010

[12] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *In Workshop on Statistical Learning in Computer Vision, ECCV*, 2004, pp. 1–22.

[13] T. Tuytelaars and K. Mikolajczyk, "Local invariant feature detectors: a survey," *Found. Trends. Comput. Graph. Vis.*, vol. 3, pp. 177–280, July 2008.

[14] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*. The MIT Press, 2001.

[15] R. Jain, *The Art of Computer Systems Performance Analysis: Techniques for Experimental Design, Measurement, Simulation, and Modeling*. John Wiley and Sons, Inc., 1991.