# Video Summarization from Spatio-Temporal Features

Robert Laganière, Raphael Bacco, Arnaud Hocevar
VIVA lab
SITE - University of Ottawa
K1N 6N5 CANADA
laganier@site.uottawa.ca

Patrick Lambert, Grégory Païs
LISTIC
Polytech'Savoie
Annecy, FRANCE
patrick.lambert@univ-savoie.fr

Bogdan E. Ionescu
LAPI
University "Politehnica" of Bucharest
061071 Bucharest, ROMANIA
bionescu@alpha.imag.pub.ro

## ABSTRACT

In this paper we present a video summarization method based on the study of spatio-temporal activity within the video. The visual activity is estimated by measuring the number of interest points, jointly obtained in the spatial and temporal domains. The proposed approach is composed of five steps. First, image features are collected using the spatio-temporal Hessian matrix. Then, these features are processed to retrieve the candidate video segments for the summary (denoted clips). Further on, two specific steps are designed to first detect the redundant clips, and second to eliminate the clapperboard images. The final step consists in the construction of the final summary which is performed by retaining the clips showing the highest level of activity. The proposed approach was tested on the BBC Rushes Summarization task within the TRECVID 2008 campaign.

## Categories and Subject Descriptors

I.2.10 [**Computing Methodologies**]: Artificial Intelligence—*vision and scene understanding*; H.3.m [**Information Systems**]: Information Storage and Retrieval—*miscellaneous*

**General Terms:** Algorithms, Perfomances.

**Keywords:** Video abstract, spatio-temporal features, Hessian-Laplace.

## 1. INTRODUCTION

The volume of digital video is continuously growing and users are requiring tools to deal with this very large amount of data. Among the different existing tools, video summarization is essential because it allows to quickly grab the relevant content of a video. In the literature [10] [8], there are a lot of papers proposing efficient approaches to video summarization. All these approaches differ according to the form of the abstract (still-image - collection of salient images - or video skim - collection of video segments), to the information sources (internal - provided by the video stream- or external), to the video modality handled (image, sound or text) and to the features extracted for each modality. Generally, the main problem with all these techniques is the gap between the information retrieved from video data and the semantic concepts required to achieve an efficient summary.

In this paper, we try to overcome this issue by addressing the video skimming task using activity-based features. As the aim is to get a very short summary from video rushes, i.e. less than 2%, we have decided to measure the activity within the video using spatio-temporal features. Therefore, we exploit the hypothesis that, for our task, the relevant candidate video segments are the ones containing high activity level. Considering previous works, the originality of this approach is in the joint detection of spatial and temporal features.

The layout of this article is as follows. In the next section, we give a global presentation of the proposed approach. Techniques for getting spatio-temporal features are discussed in Section 3 and the way these features are used to get the clips (the video segments candidate for the summary) is detailed in Section 4. Section 5 presents the algorithm for constituting the final summary. It includes two post-processing steps: redundancy reduction and clapperboard detection (presented in Section 6). Some results are proposed in section 7. Section 8 concludes this article.

## 2. THE PROPOSED APPROACH

The proposed approach is described with Figure 1. It consists of five processing steps. First the spatio-temporal features are extracted. This extraction is based on the use of the spatio-temporal Hessian matrix and provides a measure of the activity within each frame. Then, this information is processed to retrieve the keyframes and thereafter the video segments (denoted as 'clips' in the following) which form the candidate set used to build the summary. The basic principle relies on the selection of the segments where activity level is high. As the video rushes contain a lot of redundancy and junk frames (e.g. clapperboard frames), two specific steps are designed to first detect redundant clips and second to eliminate clapperboard images. The final step consists in fusing together of all these pieces of information to achieve the final summary taking into account the time constraint.

## 3. SPATIO-TEMPORAL FEATURES

Image features have been largely used in computer vision to perform matching, tracking and recognition tasks. These features, or interest points, are generally detected as local


*TVS'08,* October 31, 2008, Vancouver, British Columbia, Canada.

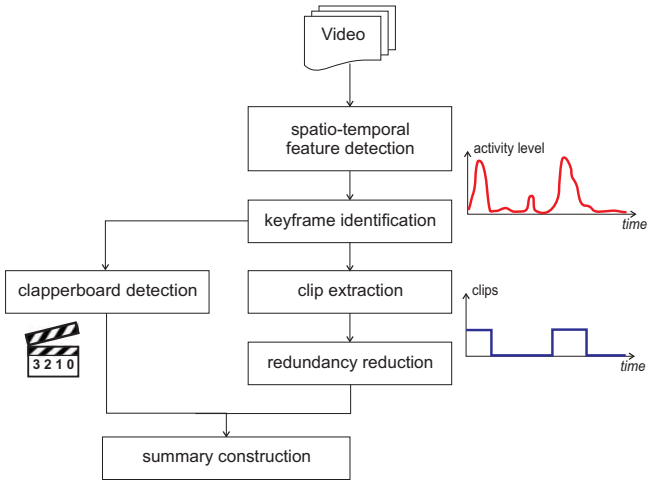


**Figure 1: Diagram of the proposed approach.**

image structures exhibiting significant intensity variations in more than one direction. Since the image partial spatial derivatives are, most often, estimated using Gaussian filters, the idea of performing interest point detection at different scales rapidly emerge as a powerful framework for scale-invariant feature detection. Lindeberg [4] has proposed a scale-invariant detector where a feature is declared at local maxima of the normalized Laplacian in scale-space. Lowe [5] proposed to approximate Laplacian using DoG filters. Mikolajczyk and Schmid [6] introduced the scale-adapted variant of the classical Harris operator. Other operators have also been proposed, some of them are compared, see [7].

More recently, spatial interest point operators have been extended to the third temporal dimension, i.e. the concept of spatio-temporal feature detectors have been introduced. Motivated by the success of feature-based object recognition, visual features in space-time have been used in event description and recognition. The interest points in a video sequence then become the ones with large variations of pixel intensities in both spatial and temporal dimensions. Laptev and Lindeberg [3] used the idea of Harris points and constructed a $3 \times 3$ spatio-temporal second-moment matrix. Like in the Harris operator, this matrix is composed of second order Gaussian derivatives averaged in a predefined neighborhood. The spatio-temporal interest points are the ones for which this matrix has 3 significant eigenvalues. They applied these features to the context of video interpretation. These features have also been used for video synchronization [2].

In this paper, we use spatio-temporal features in the context of video summarization. To this end, we based our operator on the Hessian matrix that has been shown to produce stable features performing well in object recognition and feature matching applications [6] [7] [1].

The Hessian matrix of a spatio-temporal signal $I(x, y, t)$ is given by:

$$H(I) = \begin{bmatrix} \frac{\partial^2 I}{\partial x^2} & \frac{\partial^2 I}{\partial x \partial y} & \frac{\partial^2 I}{\partial x \partial t} \\ \frac{\partial^2 I}{\partial x \partial y} & \frac{\partial^2 I}{\partial y^2} & \frac{\partial^2 I}{\partial y \partial t} \\ \frac{\partial^2 I}{\partial x \partial t} & \frac{\partial^2 I}{\partial y \partial t} & \frac{\partial^2 I}{\partial t^2} \end{bmatrix} \quad (1)$$

where $\frac{\partial}{\partial x}$ denotes the partial derivative with respect to $x$. In object recognition, a scale-adapted version of this matrix is generally used. The derivatives are then estimated by convoluting the image sequence with appropriate Gaussian filters. In the case of a three-dimensional signal $(x, y, t)$, 3D Gaussian filters are, in principle, required which would be computationally expensive to applied. However, thanks to the separability property of these filters, it is possible to apply the temporal and spatial components separately. The term $\frac{\partial^2 I}{\partial x \partial t}$ is then approximated by:

$$\frac{\partial g_{\sigma_s}(x, y)}{\partial x} \otimes \left( \frac{\partial g_{\sigma_t}(t)}{\partial t} \otimes I(x, y, t) \right) \quad (2)$$

where $g_{\sigma_s}(x, y)$ is the 2-dimensional Gaussian with variance $\sigma_s^2$ and $g_{\sigma_t}(t)$ is the 1-dimensional Gaussian with variance $\sigma_t^2$; $\otimes$ denotes the convolution operator. The other terms of the matrix are computed similarly. The variance of the Gaussian filters controls the spatial and temporal scales at which the derivatives are evaluated. In our experiment we used a fixed value of 1.5 for both $\sigma_t^2$ and $\sigma_s^2$.

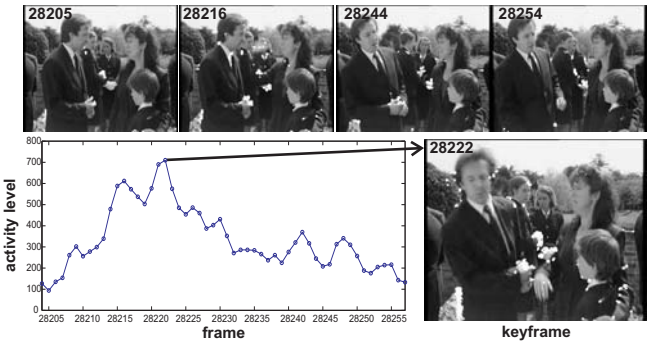


**Figure 2: Representing the activity level (number of detected features) vs. frame number for a segment.**

A spatio-temporal feature point is declared whenever the determinant of the Hessian matrix, $det(H)$, exceeds a predefined threshold. It is observed that such features occur at prominent motion of visual interest points. Figure 2 shows the detected features in few frames of a short video segment. The next section describes how these visual features are used to extract salient activity clips in a video sequence.

## 4. CLIP EXTRACTION

Our objective is to summarize a video sequence by means of extracting only the clips inside the sequence in which the level of activity is significantly high. These clips of salient activity can correspond to scenes where characters are performing quick actions, or where multiple entities are interacting together; they most often constitute 'key scenes' that summarize well the action within the movie. Our main observation, on which the approach presented in this paper is based, is that salient activities in a video sequence will generate important spatio-temporal variations of the pixel intensity. Consequently, it seems therefore appropriate to use the density of spatio-temporal features to detect 'key clips'.

We define a clip, $c_{[t_o, t_f]}$ , as a short video segment (subsequence) which is composed of the sequence starting at frame $I(t_o)$ and ending at frame $I(t_f)$. Our strategy consists then in extracting a set of disjoint clips $\mathcal{C}$ from the original

video which are found to be representative of the film content. These clips will stand as the basis for generating the final video summary.

To be considered significant, a clip must encompass a scene exhibiting high activity level. In our approach, the candidate clips will be the ones inside which the level of activity reaches a local maximum. The keyframe associated with a given clip will be therefore the one with the highest value of $n(t)$. It follows from this definition that one can identify the candidate set of clips simply by locating the local temporal maxima of $n(t)$. Figure 2 illustrates how the number of feature point $n(t)$ evolves with the action in a video segment.

First we extract the Hessian spatio-temporal features for each frame $I(t)$, with $t = 1, ..., T$ and $T$ representing the video sequence length. The number of detected features, $n(t)$, in each frame is counted. The clip selection process then starts with an empty set of candidate clips, $\mathcal{C} = \emptyset$. The clips will then be iteratively added to $\mathcal{C}$ by considering, at each iteration, the frame with the highest feature count, that is:

$$t_{max} = \arg\max_t \{n(t)/I(t) \notin \mathcal{C}\} \quad (3)$$

where $I(t) \in \mathcal{C}$ if $\exists\ c_{[t_o,t_f]} \in \mathcal{C}$ with $t \in [t_o, t_f]$. Each of the thus extracted frames constitutes a keyframe of the film. In order to attenuate the effect of local variations, we first smooth out the activity level curve by applying a mean filter (of width equals to 11).

Once a keyframe identified, the next step consists in determining the boundaries of the enclosing clip (i.e. finding the $t_o$ and $t_f$ of a clip). This is done by identifying the interval inside which the level of activity remains significant enough. Starting from the current $I(t_{max})$, we seek forward in the sequence to find the first frame for which the number of spatio-temporal features is less than a fraction $\alpha$ of $n(t_{max})$, that is $I(t_f)$ with:

$$t_f = \min_t \{t/n(t) < \alpha \cdot n(t_{max})\} \quad (4)$$

where $t \in [t_{max} + 1, T]$ and $0 < \alpha < 1$ corresponds to the percentage of activity reduction (feature count) used to delimitate the clip (we used 0.2). The smaller is the value of $\alpha$, the longer will be the extracted clip. If a frame $I(t) \in \mathcal{C}$ is found before this condition is met, then $t_f$ will be the frame preceding this frame.

Similarly to the preceding step, starting from the frame $I(t_{max})$, we seek backwards in the sequence for a frame such that:

$$t_o = \max_t \{t/n(t) < \alpha \cdot n(t_{max})\} \quad (5)$$

where $t \in [1, t_{max} - 1]$.

Once the clip boundaries identified, we can then estimate the level of activity in a clip $c_{[t_o,t_f]}$ as:

$$H(c_{[t_o,t_f]}) = \sum_{t=t_o}^{t_f} \log n(t) \quad (6)$$

The number of feature points is used here to measure the magnitude of the activity; using the *log* function allows to avoid that a single isolated frame with high activity impacts too much on the global activity level of the clip. This clip activity level measure will be used to build the summary, as explained in the next section.

Note also that with this approach, the extracted clips will vary in duration; the method aims at encompassing the action from the point where is starts until it ends. But, in some cases, one might want to restrict the duration of each clip to a certain max, $d_{max}$. This can be done using the following approach: if the extracted clip exceeds this maximum, the clip is cropped on each side of its representative keyframe by increasing the value of the threshold $\alpha$ until we achieve the required clip duration.

After determining a clip boundaries, the clip $c_{[t_o,t_f]}$ is included in the clip set $\mathcal{C}$ and the process starts over with the selection of a new keyframe. In principle, one can continue to extract keyframes until no more frames are available, since the subset of clips from $\mathcal{C}$ that will form the summary will be selected at the final step. However, in practice, it is useless to continue clip extraction when the number of features in the current keyframe becomes very low.

At this point we have a set of candidate clips extracted from the entire video sequence. We now want to select a subset of these clips which when assembled together gives rise to a summary of a user-specified duration $SD$. This is explained in the next section. However, a first prune out of this set can be achieved simply by eliminating the clips that are too short. Indeed, too short clips are visually uninteresting and usually not very relevant. The duration of a clip being simply given by:

$$d(c) = t_f - t_o \quad (7)$$

We remove any clip such that $d(c) < d_{min}$ (e.g. $< 1$ sec.).

# 5. BUILDING THE SUMMARY

The summary will be composed based on the following considerations: first, the clips with higher level of activities are generally more meaningful and should then be considered in priority; second, a summary should present a good variety of visual elements, consequently only one instance of a subset of clips with very similar content should be included in the final summary; finally, to ensure that the summary is complete and representative of the original movie, the extracted clips should be well distributed over the temporal duration of the film.

These observations suggest that the summary can be built through an incremental procedure. Starting with a new set of clips, $\mathcal{S}$, initially empty, clips from $\mathcal{C}$ are transferred to $\mathcal{S}$ until no more clips can be added without exceeding the preestablished maximal summary duration, $SD$. The clips are added to the summary set according to the following rules:

1. Extract the clip $c_{max} \in \mathcal{C}$ with the highest activity level, $H(c_{max}) > H(c), \forall c \in \mathcal{C}$ with $c_{max} \neq c$.
2. If $[d(c_{max}) + \sum_{c \in \mathcal{S}} d(c)] > SD$ then $c_{max}$ is deleted.
3. If $\exists\ c_s \in \mathcal{S}$ such that $D_v(c_s, c_{max}) < DV_{low}$ then $c_{max}$ is deleted, where $D_v()$ measures the visual distance between clips and $DV_{low}$ is a threshold under which clips are judged very similar in content.
4. If $\exists\ c_t \in \mathcal{S}$ such that $D_t(c_t, c_{max}) < DT_{min}$ AND $D_v(c_t, c_{max}) < DV_{high}$ then $c_{max}$ is deleted. $D_t()$ measures the temporal distance between two clips that must be greater than the threshold $DT_{min}$. When the visual distance between two clips exceeds the threshold $DV_{high}$ then the two clips are considered very dissimilar in content.

5. If the conditions 2., 3. and 4. are not fulfilled then the current clip $c_{max}$ is added to $\mathcal{S}$.

The visual distance between clips, $D_v()$, is computed using a simple and yet efficient method which is based on the computation of color histograms. To capture the visual global color signature of a clip, a mean color histogram is computed on a percentage $p\%$ of the clip frames (usually $p \in [15; 30]\%$). The retained frames are color reduced using an accurate color reduction scheme, the Floyd-Steinberg dithering algorithm. The mean histogram is computed as:

$$\bar{h}_c(i) = \frac{1}{N_c} \sum_{j=1}^{N_c} h_c^j(i) \quad (8)$$

where $N_c$ is the total number of retained frames in the clip $c$, $h_c^j()$ is the color histogram of the frame $j$ from the clip $c$ and $i$ is the color index from the reduced color palette.

The visual similarity function between two clips $c_1$ and $c_2$, $D_v(c_1, c_2)$, is given then by the Euclidian distance between their histograms, thus $D_v(c_1, c_2) = D_E(\bar{h}_{c_1}, \bar{h}_{c_2})$.

Therefore, *Step 3* of the algorithm assures the exclusion of clips that are visually very similar to the ones already selected, while *Step 4* ensures that the clips from the summary provide an adequate temporal coverage of the original film. However, in order to avoid to exclude significant video segments, clips that show a high level of dissimilarity will be accepted even if they are in close temporal proximity.

This procedure is repeated until the set $\mathcal{C}$ becomes empty. Finally, the clips in $\mathcal{S}$ are reordered with the respect to time and then assembled together to produce the summary.

## 6. DETECTING THE CLAPPERBOARDS

In particular cases, such as for video rushes, the original film often contains numerous occurrences of meaningless visual elements. Color bars and clapperboards are the two most common such visual elements. From the point of view of the summarization task, these elements are not significant and must be disregarded.

In our approach, color bars are easily excluded as they do not contain motion. On the other hand, clapperboard scenes always involve significant motion and they are selected as clips of salient activity by our method. In fact, through our experimental tests, it turns out that our approach acts as an excellent clapperboard detector. Indeed, the exact instant in time where the clapperboard is clapped generally corresponds to a peaks of activity on the spatio-temporal features density graph; this means that the extracted keyframe in a clapperboard shot will be the frame showing the clapperboard clapping, which generally occurs when the clapper is located right at the center of the frame. In addition, clappers have a very specific visual pattern that can be easily characterized. In consequence, it becomes relatively simple to reliably identify clapperboards in the keyframes extracted by our algorithm just by looking at the color histogram of the center region of the image.

Clapperboards are either white with black edges or black with white edges. Consequently, a clapperboard will produce an histogram with a majority of black and white pixels. We therefore first look for unsaturated pixels; the saturation (or colorfulness) being measured as:

$$Sat(R, G, B) = max(R, G, B) - min(R, G, B) \quad (9)$$

if this value is less than a threshold (we used 64), then it is considered unsaturated (a gray level). If the majority of the pixels inside a center window (half the size of the image) are unsaturated then we count the number of black and white pixels. For a white clapperboard, the number of bright pixels must be sufficiently large while the number of dark pixels must be low (the intensity being determined by the sum $R + G + B$). This simple algorithm works surprisingly well as it can be seen in Figure 4.


**Figure 4: Several keyframes extracted following the method from Section 4 for one of the TRECVID08 video rush (red X's=detected clapperboards).**

## 7. EXPERIMENTAL RESULTS

The result of the proposed summarization method is illustrated in Figure 5 where a 34 sec. summary was produced from a 30-min video rush. The graph shows the evolution of activity level (number of detected spatio-temporal features) across the entire sequence. Note that the very high peaks mostly correspond to the shot boundaries in the video. Indeed, scene cuts produce instantaneous spatio-temporal changes when transiting from one shot to another. However, these peaks are discarded from the summary because of their very short duration and of the very low total activity level of their corresponding clips. The extracted keyframes, shown in Figure 4, are located on the graph using dots. The video segments that form the summary built from the process described in Section 5 is represented by the set of thick lines. Nine clips were used to build this summary; they correspond to the keyframes 2, 4, 6, 10, 17, 18, 30 and 31 in Figure 4. Inclusion of duplicate scenes such as the ones represented by keyframes 10 and 12 or by keyframes 30, 33, 36 have been avoided through the similarity measure described in Section 5.

In the framework of TRECVID 2008 [9], our produced summaries were judged to have a very pleasant rhythm (rank 5 on 43) while in terms of junk and duplicate scenes, the method obtained average results (rank 24 for junks and 12 for duplicates). However, a low 26% was obtained for the inclusion criteria (rank 39). The fact that the video segments extracted by our method are centered at the peak

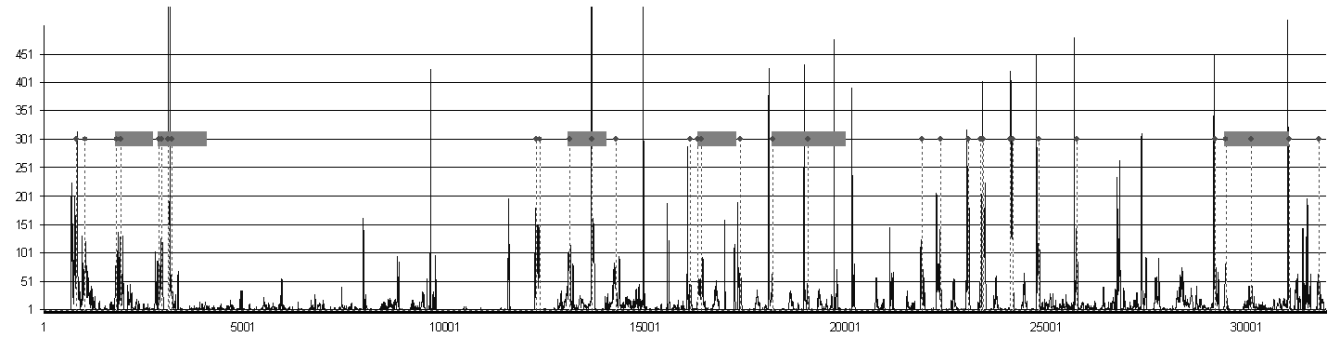


**Figure 3: Graphical representation of $n(t)$ vs. frame number for a TRECVID video. The keyframes shown in Figure 4 are located with dots while the frames included in the summary are represented by gray thick lines.**

of activity level constitutes an important factor in the production of summaries with pleasant tempo. It ensures that each clip is extracted with an appropriate time window that includes the starting and the ending of the action shown. Also, the fact that we selected a maximal clip duration of 3 seconds participates to the production of summaries with good fluidity. However, individual clips of longer duration reduces the total number of clips that can be included in the summary and hence reduces the probability of inclusion. If, for an equivalent summary duration, one wants to include more clips in the summary then, with our method, the maximum clip duration should be reduced. Each clip will then be shorter and the procedure will thus include a larger number of clips. However, this will result in a saccadic summary.

# 8. DISCUSSIONS AND CONCLUSION

This paper presented a method for video summarization based on the detection of spatio-temporal features. Summaries were produced based on the assumption that the relevant candidate video segments that should be selected to build a summary are the ones exhibiting high activity levels. This is certainly a debatable statement but the subjective testing of the results seems to confirm its validity. High activity video segments generally correspond to clips where characters are performing important actions. Some important scenes can however be missed such as panorama panning or shootings of inactive persons; such scenes are of lesser importance when summarizing the story of a film, but they might be relevant when summarizing a collection of rushes. It also happens that some background motions (e.g. water motion) produce scene with high activity levels; these are then undesirably included in the summary.

The detection of duplicate scenes is based on the comparisons of color histograms. Currently, the thresholds used in these comparisons are set empirically. We currently work on an adaptive technique based on clip clustering that will allow to automatically determine the value of these parameters. The detection of the spatio-temporal features also requires some parameters to be set. However, the method is quite tolerant to these as it relies on the detection of local maxima; similar results being obtained for a large range of values. The most critical parameters are the minimum and maximum clip duration as well as the summary duration; these are however entirely subjective and depends on the user's objectives.

# 9. ACKNOWLEDGMENTS

This work was partially supported by the Rhône-Alpes region, Research Cluster 2 - LIMA project and by CNCSIS - National University Research Council of Romania, grant 6/01-10-2007/RP-2. Part of this work has been done while the first author was invited professor at Polytech'Savoie (Université de Savoie).

# 10. REFERENCES

[1] A. Bhatia, R. Laganiere, G. Roth, Performance Evaluation of Scale-Interpolated Hessian-Laplace and Haar Descriptors for Feature Matching, International Conference on Image Analysis and Processing, pages 61-66, Modena, Italy, 2007.

[2] D. Wedge, D. Huynh, P. Kovesi, Using Space-Time Interest Points for Video Sequence Synchronization, IAPR Conference on Machine Vision Applications, pages 190-194, Tokyo, Japan, 2007.

[3] I. Laptev, T. Lindeberg, Space-time interest points, ICCV03, pages 432-439, 2003.

[4] T. Lindeberg, Feature Detection with Automatic Scale Selection, International Journal of Computer Vision, 30(2), pages 79-116, 1998.

[5] D. Lowe, Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision, 20, pages 91-110, 2003.

[6] K. Mikolajczyk, C. Schmid, Scale and affine invariant interest point detectors, International Journal of Computer Vision, 60(1), pages 63-86, 2004.

[7] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, L. Van Gool, A Comparison of Affine Region Detectors, International Journal of Computer Vision, 65(1-2), pages 43-72, 2005.

[8] A.G. Money, H. Agius, Video summarisation: A conceptual framework and survey of the state of the art, Journal of Visual Communication and Image Representation, 19(2), pages 121-143, 2008.

[9] P. Over, A.F. Smeaton, G. Awad, Trecvid 2008 BBC Rushes Summarization Evaluation, Proceedings of the ACM International Workshop on TRECVID Video Summarization, pages 1-20, Vancouver, Canada, 2008.

[10] Ba Tu Truong, S. Venkatesh, Video abstraction: A systematic review and classification, ACM TOMCCAP, 3(1), 2007.