# Video object mining: issues and perspectives

Jonathan Weber, Sébastien Lefèvre, Pierre Gancarski

# Video Object Mining : Issues and Perspectives

Jonathan Weber
and Sébastien Lefèvre
and Pierre Gançarski
Image Sciences, Computer Sciences and Remote Sensing Laboratory (LSIIT)
University of Strasbourg – CNRS
Pôle API, Blvd Sébastien Brant, PO Box 10413, 67412 Illkirch Cedex, France
Email : {j.weber,lefevre,gancarski}@unistra.fr

***Abstract*—Today, video is becoming one of the primary sources of information. Current video mining systems face the problem of the semantic gap (i.e., the difference between the semantic meaning of video contents and the digital information encoded within the video files). This gap can be bridged by relying on the real objects present in videos because of the semantic meaning of objects. But video object mining needs some semantics, both in the object extraction step and in the object mining step. We think that the introduction of semantics during these steps can be ensured by user interaction. We then propose a generic framework to deal with video object mining.**

## I. Introduction

After the massive increase of text and more recently image data available on databases and on the world wide web, we observe today an expansion in the amount of video data. Video is becoming one of the primary sources of information (e.g., YouTube [1] serves up more than 100 million videos a day online). The temporal aspect of videos prevents the efficient browsing of these very large databases. However, apart for temporal segmentation where it plays an almost exclusive role [2], the temporal dimension is of limited use in existing algorithms related to video mining. Video mining [3] is the extension of data mining to the video domain. Commonly, data mining [4] is considered as the process of extracting information/knowledge from large amounts of data. A video is a temporal image sequence, eventually coupled with audio data. In this article however we only focus on image sequences. According to these definitions, a Video Mining System (VMS) is a system which is able to extract information from a large repository of image sequences.

Several studies have already surveyed the field of video mining. Mainly, each one reviews a subfield of the video mining domain such as video indexing or video summarization but not the whole video mining field. Among the earliest works, Idris and Panchanathan [5] present some video indexing methods and point out that the natural level for analyzing visual content should be the object level. Brunelli *et al.* [6] also study video indexing systems and notice that, in 1999, detecting generic objects could not be achieved with state-of-the-art methods. Ten years later, Brezeale and Cook [7] discuss the level considered in the classification task: while most of the reviewed papers proposed to work on the global video level, only few deal with the shot or the scene level. None are related to the object level. Money and Agius [8] survey the video summarization task. They propose focusing on user-based information to compute personalized summaries and to summarize the video content at a higher semantic level. Ren *et al.* [9] explore the use of spatio-temporal information for video retrieval. They point out the effectiveness of knowing the spatio-temporal relations between the objects for video retrieval. Contrary to these approaches, we do not focus on one specific task of video mining but consider here the whole video mining context. In agreement with conclusion of [5], we suggest performing object-oriented video mining. Thus, in this paper, we precisely focus on the role of the object in the video mining task and discuss how to set the object as the fundamental element of the video mining process.

In this article, we first introduce a new taxonomy to characterize the different video mining systems. Then we focus on the current role played by objects in video mining systems and review recent works using the proposed taxonomy. Subsequently, we discuss deeper issues related to video object mining. We determine the characteristics of a video object mining system, present the problem of VO extraction and study how to introduce semantics in VOMS. Finally we propose a generic scheme for a framework for designing VOMS.

## II. Characteristics of video mining systems

Many aspects have to be taken into account when designing a new system for video mining. In this section, we identify such aspects and we introduce some terms which can be used to characterize VMS.

### A. Tasks of VMS

All VMSs do not have the same purpose due to the different needs of their users. Video repositories need huge storage capacities and manual video mining is time expensive, so related research works focus on systems to automatically perform the task usually performed by a human user or on systems that alleviate user intervention. *Video summarization* (Sum) aims to provide short and meaningful representations of videos in order to allow the users to retrieve their topics and contents without watching the entire videos. *Video indexing* (Ind) is the process of characterizing a video in order to be able to retrieve it quickly afterwards using specific queries .

*Video classification* (Cla) aims to group together videos with similar content and to disjoin videos with non-similar content. *Content-based video retrieval* (Ret) allows users to retrieve videos similar to a specific given video .

### B. Properties of VMS

VMSs have several properties depending on the data types (compressed or not, specific or not), on the type of information to be processed (video, scene or object), on the scales which are used to compute features, on the features to be extracted and on the role of the user in the system. In this section, we introduce and precisely describe these different aspects. Let us observe that, to our best knowledge, there is no existing VMS generic enough to deal with all these aspects.

*Data:* A VMS may deal with various types of video data. A video can be uncompressed (U), or it can be compressed (C) with different algorithms . Compression ensures a more efficient storage but requires a decompression process before visualization and possibly induces data loss which may be an issue in certain fields . Dealing with compressed video is faster because less data is processed but extraction of visual concepts may be hardly achieved, while analysis of visual content is straightforward from uncompressed videos (with a higher computational cost though). VMS can also be dedicated to a specific (S) kind of video (e.g., sport broadcasts or news) or to generic (G) videos. Processing specific videos can achieve better results because domain knowledge can be involved while processing generic videos allows VMS to deal with any possible video repository.

*Elements:* Whichever tasks a given VMS is dedicated to, it can be applied to different elements, from the entire video to a single pixel. The first step in video mining is generally to extract the specific elements to be analyzed. *Entire video* (Vid) is the classical element to be processed and the easiest one because it is already available in the appropriate form. Nevertheless, it can be meaningless if the video contains too many unsimilar scenes. A *scene* (Sce) is composed of shots in the same context (time, space) and for this reason, may be difficult to be extracted. *Shot* (Sho) is the video segment delimited by two abrupt or progressive transitions, its extraction is a widely studied problem and many methods have been proposed to solve it [10]. *Frames* (Fra) are the elementary spatial-only units of video as a video is a sequence of frames. *Object* (Obj) is according to us the most meaningful element but its use is limited by the intrinsic difficulty to extract real object (e.g., car) and its evolution with time. *Region* (Reg) is an object but does not rely on any semantic meaning, it is just a connected set of pixels. Finally, *pixel* (Pix) is the smallest element but it has no meaning if taken alone. Figure 1 summarizes all these notions.

*Scales:* To analyze video and characterize these elements, some features are commonly extracted from the data. These features can be computed at different scales . Scale is linked to the element under consideration and the features which are used. With scale *global* (Glo), video features are related to the entire image sequence . Scale *block* (Blo) divides the

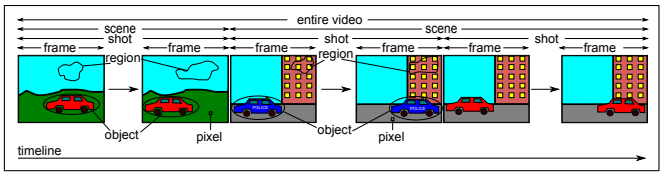


Fig. 1. The different VMS elements.

video into blocks using a spatial grid and video features are computed on each block independently. Contrary to block scale, scale *region* (Reg) does not divide the video into blocks but into regions of various shapes through a segmentation step. Video features are computed on each region independently. Scale *Object* (Obj) is nearly the same as region scale, but objects are assumed to bring a semantic meaning. Scale *Points of interest* (POI) relies on pixels with a neighboring configuration that is remarkable in any sense. Video features are computed on each point of interest independently. Scale *Pixel* (Pix) is the smallest possible scale in which video features are computed on each pixel independently but it seems to be meaningless. Figure 2 summarizes all these notions.

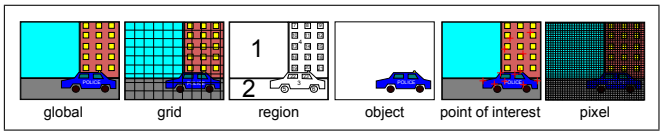


Fig. 2. The different scales VMS elements can be processed at. For the sake of readibility, only spatial information is considered in this figure (best viewed in color).

*Features:* There are many video features available to describe video content. We may describe local or global motion (Mot) but also the colors (Col). It is also possible to characterize the texture (Tex) present in the visual data. We may abstract the shape and edge (Sha) to describe the morphology of objects present in the video . Besides these classical kinds of features, there are many more specific features constantly proposed in the literature. Choosing appropriate features in a given VMS is not trivial because each feature aims to give specific information about the video content. .

### C. Role of the user in the mining process

The place of the user is a critical point in VMSs. The influence of the user may be categorized into four different levels. *Null*, when no intervention of the user is needed. *Semi-supervised* (Semi), if the user has to provide some examples or to validate/invalidate some results. . *Supervised* (Sup), when the user has to provide a complete set of (labeled) training data in order to configure the system for the dataset to process. . *Parameters* level (Param) corresponds to the case where user has to configure the different parameters of the system.

The taxonomy we have introduced in this section allows to efficiently describe and characterize a VMS. As we illus-

trate it in the next section this taxonomy is adapted for the comparaison of different VMS.

## III. Object usage in video mining

In this section, we present recent work dealing more or less with object-related video mining, to highlight the current trends in this domain. Table I summarizes all the characteristics of the reviewed systems, according to the taxonomy introduced in the Section II.

| Method | | Task | Data | Element | Feature | Scale | User |
|---|---|---|---|---|---|---|---|
| Anjulan | [11] | Ret | U,G | Obj | LIR,SIFT | Reg | Param |
| Anjulan | [12] | Cla | U,G | Obj | LIR,SIFT | Reg | Param |
| Avila | [13] | Sum | U,G | Vid | Col,LP | Glo | Param |
| Basharat | [14] | Ret | U,G | Vid | SIFT,Col,Tex,Mot | Reg | Null |
| Chevalier | [15] | Ret | C,G | Obj | RAG | Reg | Param |
| Gao | [16] | Ret | U,G | Sho | OFT | Blo | Param |
| Liu | [17] | Ret | U,G | Vid | Diverse | Obj | Param |
| Ren | [18] | Sum | U,G | Vid | PLR,ECR,HCC | Glo | Param |
| Sivic | [19] | Ret | U,G | Obj | SIFT | Reg | Param |
| Teixeira | [20] | Cla | U,S | Obj | SIFT | Obj | Sup |
| Zhai | [21] | Sum | U,G | Vid | KNNG | Glo | Param |

TABLE I
SUMMARIZATION OF CURRENT TRENDS (ONLY THE FIRST AUTHOR IS STATED)

Anjulan and Canagarajah [11] present an object video retrieval system. First, local invariant regions (LIR) are extracted and then characterized with SIFT . Then, LIRs representing the same object in successive frames are linked, forming a temporal track of similar LIRs. These tracks are clustered and the queries, region selected by the user, are compared to the cluster center.

Anjulan and Canagarajah [12] use the method they defined in [11] for extracting LIR tracks but instead of considering a retrieval task, they propose a method for clustering objects in the video. For each shot, tracks representing the same object are grouped into a cluster assuming the fact that two tracks belong to the same object if they are spatially close and their spatial distance is constant in the frames in which they both appear. Then, instances of the same object in different shots are grouped.

Avila *et al.* [13] propose a simple video summarization method. On each frame a color histogram and line profiles (LP) are computed. Frames are then clustered using K-Means algorithm according to histogram and LPs. The summary is composed of the most representative frames of each cluster. This is a static summary, motion information is not being taken into account.

Basharat *et al.* [14] define a video retrieval system based on spatio-temporal volume correspondences. Spatio-temporal volumes are 3D objects representing spatial regions and their evolution through time. First, points of interest are extracted in each frame and described by SIFT. Similar points in successive frames are linked to build trajectories. Spatio-temporal volumes are obtained from these trajectories, assuming that a region is composed by points with similar trajectories. The volumes obtained are then characterized using four types of features: interest point descriptors, color, texture and motion. Comparison with query is achieved with a bi-partite graph. The edge weights of the graph are the similarity measurements between two volumes of two different videos.

Chevalier *et al.* [15] deal with the retrieval of objects in a video. The video is first segmented by a watershed algorithm applied on a low-resolution frame (DC). This region growing algorithm produces a partition of regions corresponding to segmented objects. The objects are then represented by region adjacent graphs (RAG). A similarity measurement is performed by a graph comparison method.

Gao *et al.* [16] present a shot-based video retrieval system relying on motion analysis. Motion is characterized in three steps. The optical flow is computed first, then it is spatially divided into optical flow cubes. These cubes are used to build optical flow tensors (OFT). High dimensionality of the OFTs is reduced by linear discriminant analysis. The retrieval part of the system is based on Hidden Markov Models.

Liu and Chen [17] present a video retrieval system based on automatic extraction of objects of interest. First, regions are extracted and characterized by SIFT. These regions are then categorized into regions belonging to objects of interest and regions belonging to background. This categorization is made by an EM algorithm. Then a bounding box is built around each object of interest and the bounding box content is described with different features. The retrieval step is performed by a new ensemble-based matching method.

Ren and Zhu [18] define a method for video summarization based on machine learning. First, several features are extracted such as pixel likelihood ratio, edge change ratio and histogram correlation coefficient. A neural network is used to detect transitions between shots using previous features. Representative frames are then extracted according to changes in edge, color and texture content. Summary of the video is defined by all the extracted representative frames.

Sivic and Zisserman [19] transpose the principles of text-based search to video object retrieval. In each frame, regions are extracted and represented by SIFT. These regions are tracked through the video. After a filtering step, remaining regions from a subset of the video are clustered creating a visual vocabulary. Most common regions are rejected and an inverted file indexing structure is built. User query is a region from a video which is analyzed to extract the visual words it contains. Then, results are retrieved using visual words frequency and spatial consistency.

Teixeira and Corte-Real [20] propose a system for classifying object in videos and apply it to visual surveillance. First, objects are segmented and tracked, then their descriptions are made by SIFT. Descriptions are quantized using bag-of-visterms based on a visual vocabulary tree. Objects are then classified using Learn++.MT algorithm.

Zhai *et al.* [21] elaborate a video summarization method based on graph clustering. A k-nearest neighbor graph (KNNG) of the video frames is first constructed. This graph is partitioned in connected components representing clusters of similar frames. If the connected components contain more nodes than a given threshold, an ISOMAP method is applied, followed by a mixture model clustering. One frame per cluster

is selected to represent the cluster in the summary. As in [13], motion information is not taken into account.

These recent articles (from 2007 to 2009) work almost all with generic data and mainly with uncompressed data . Concerning the tasks, the main trend seems to be the retrieval. This is not surprising as the first thing an user wants to be able to do is to retrieve the data she/he needs, especially in case of high quantity of data. Moreover we notice that summarization is also an important field of research, in fact the possibility of knowing the content of a video without watching the entire video is a great gain of time. Except for the summarization task, the classical element starts to be the object. However the object scale is far from being the usual scale, global and region scales are still the most common one since automatic semantic segmentation can be hardly achieved. There are a lot of different features used but SIFT descriptors are used in several methods and seem to be the more efficient descriptors at present time. About the role of the user, we observe that, except for Basharat *et al.* [14], all the methods ask user for parameters, from easy to tricky setting and one (Teixeira and Corte-Real [20]) is supervised.

## IV. Towards object-oriented video mining

The study of the current trends in video mining shows that if the object or the region scales seem to be commonly adopted, the usual levels of analysis are still the video or the shot levels except for the particular context of the retrieval where there are methods which directly deal with objects. In video analysis, most of the information is brought by VOs and their temporal evolution. Some semantic information can be added using VO context, for instance background or other adjacent VOs. In the same way, spatio-temporal relations between VOs or non-video information such as caption text or an audio channel (not considered here) may also be used. According to these facts, performing video mining with objects as elements is relevant. Nevertheless, we face a new problem related to the VO extraction from the videos. Moreover, dealing with VOs as elements makes it necessary to take into account the semantic meaning of VOs. So there is a need to introduce semantics to the mining process. In this section we show how the choice of VOs as elements impacts the VMS characteristics.

### A. Characteristics of VOMS

Setting VO as the element of a VMS has an important influence on the other VMS characteristics (except for the tasks because all tasks can be a priori reached with VO as elements).

*Data:* The impact on data types is limited. Even if it is not trivial to extract VOs from compressed streams, there are methods which deal with this problem such as those proposed by Babu *et al.* [22], Toreyin *et al.* [23] or Hsu *et al.* [24]. Using objects as elements is possible whatever the video type. But intuitively it seems that processing specific videos is easier with objects such as elements due to the limitation of VOs diversity in a specific context. On the contrary, dealing with objects as elements makes harder the processing of generic videos due to the necessity of being able to segment all possible types of VOs.

*Scale:* Scale is greatly influenced by the choice of VO as element. Describing a VO or its content means that the highest possible scale is the object except if contextual information is considered. So, using VO as an element leads us to consider two types of scales, feature and context scales. Possible feature scales are the scales lower than or equal to the object scale while possible context scales are those higher than the object scale. Figure 3 illustrates that duality of scales. Both extracted frames shows the same object, the *Discovery* space shuttle. In the case the user has many space shuttle videos, he may desire to classify such videos according to specific situation. Describing only the object cannnot let him to achieve such a purpose. As a consequence there is a need for contextual information to distinguish the same object in different situations. Here, a user may want to make the difference between *Discovery* at its launch pad and *Discovery* in flight.

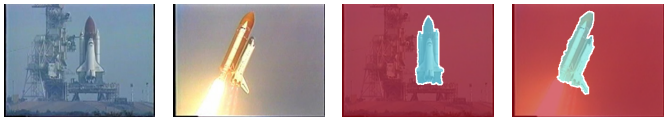

Fig. 3. Two frames extracted from *STS-53 Launch and Landing, segment 02 of 5* sequence from the Open Video Project [25] (left) and their respective segmentation (right) into the *Discovery* space shuttle (blue) and contextual information (red).

*Feature:* Every feature is able to be used in the VO-based approach but not in the same way as other elements. Global features can be applied but only to the VO and not to the whole image (except for context scale). Motion features can be applied twice, one for describing global VO movements and one for describing internal VO movements in case of complex VOs. More generally, the features used in VOMS have to be semantically discriminative.

*Role of the user:* The last (but not least) characteristic is the role of the user in the VOMS. This role of is predominant because of the semantic meaning associated with the VO concept. A fully automated system is not able to mine semantically VOs, it needs some user knowledge. Moreover, the perception of a VO can be different from one user to another according to the user subjectivity. Indeed each user expects to obtain personalized results from the VOMS. This is why a user has to be deeply involved in the video mining process to guide it. Even if this intervention is fundamental for the system, it must be intuitive and limited for the user in order to be efficient and not time expensive. This can be achieved using the paradigm of the relevance feedback as discussed in the Section IV-C. The key step of a VOMS is precisely the VO extraction from video data which is discussed next.

### B. Video-object extraction

To extract VOs from a video, every method integrate a segmentation step (except for object-oriented compressed videos and method based on interest points only). While in the VMS context, video segmentation is most of the time related to

shots [10], in VOMS the extraction step has to produce VOs as well as their evolutions over time. As underlined in Section I, the main difficulty is to bridge the gap between the raw data and the VOs. This extraction can be achieved either with a video encoding or from a segmentation.

Considering a video in a 3D space (X,Y,T), a spatio-temporal segmentation is a partition of this space into volumes, each of them representing a spatio-temporal object (i.e., spatial definition of VO and its temporal evolution). A VO has a semantic meaning. So, there is a need for semantic video segmentation methods adapted to these datasets. More generally, the introduction of semantics in a VOMS is important and is studied in the next section.

### C. Introducing semantics

Dealing with VOs in VMS requires the introduction of semantics both in the segmentation step and the dedicated analysis step (e.g., classification, indexing, etc). Low-level features, presented in Section II-B, give numeric discriminative descriptions but are not able to give semantic descriptions as human perception does. This is the problem of the semantic gap discussed previously. Answering the question "How to bridge this gap ?" leads to the question "How to introduce human knowledge in VOMS ?".

The ideal way for introducing semantics in VOMS is probably to provide samples for all possible VOs but it is clearly impossible. Another way to introduce semantics is the user relevance feedback [26]. At the end of the mining process, the user evaluates a sample of results and may correct some of them. Then the process restarts in order to take into account the user's indications. This iterative process is less time expensive for the user than the production of samples in a supervised way. It also guarantees that the result is user personalized. Moreover, relevance feedback can also be applied to the segmentation process in order to improve it. The idea is the following: the better the segmentation is, the lighter the mining step will be. Finally, determining more semantic features to describe VOs could also be a solution but it is still an open problem.

### D. Video Object Mining Framework (VOMF)

According to the previous prospects, we propose in this section a generic scheme for a Video Object Mining Framework (VOMF).

The process proposed in VOMF is shown in Figure 4 and starts with VO extraction from a video of the repository. The resulting segmentation is analyzed by the user in a relevance feedback process. If segmentation is approved, it is sent to the video mining process , otherwise segmentation errors are pointed out by the user and thus semantic information is added. A new segmentation process is applied using current segmentation, corrections and information added by the user. This cycle is repeated until the user is satisfied by the segmentation. The processing step of video mining depends on the mining task. Results of the video mining process is also analyzed by the user through the relevance feedback process.

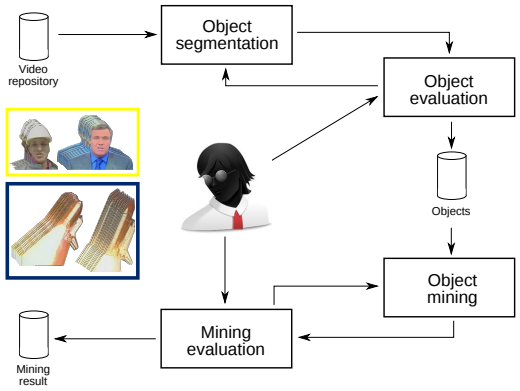


Fig. 4. VOMF: Video Object Mining Framework

If the results are satisfying, the system goes on, otherwise, as for the segmentation feedback step, the user corrects the mining results. Then the video mining process restarts and relies on these corrections and new information. As we can see, user is at the center of the process. He/she supervises the video mining process through relevance feedback and introduces semantics by correcting inaccurate results and by adding simple information, types of information depending on the objectives of the system. In order to be useful, the relevance feedback does not have to be exhaustive and should not be too time consuming. It just has to add some corrections in order to *guide* the process. This is why few corrections have to influence deeply the segmentation and the mining processes. This is developed in the next section.

### E. A user feedback to evaluate and guide video mining

VOMF has two major steps, the VO extraction and the VO mining. Each has a dedicated user feedback process to evaluate and guide the processing.

The VO extraction is, as mentionned previously, the critical step of VOMF. In fact, without correctly extracted VOs there is no possible mining. Extracting VOs in all possible kind of videos is not a trivial task. Simple user feedback is a very straight evaluation, the system presents VOs to the user and asks him/her if the extracted VOs are real objects (or in other words, the objects sought by the user). This feedback is simple but very time consuming if the user must validate all the extracted VOs and one can worry what the system shall do if the VOs are not good enough and do not satisfy the user ? Our idea takes this into account. The system presents samples of video segmentation. So, the user has three possibilities. He/she can validate the segmentation if the presented segmentation matches the user desire. He/she can correct it by a graphical interface we explain next. Or he/she can even reject it if the segmentation does not fit at all the user expectations. The correction interface presents the current segmentation and the frontiers of the initial oversegmentation. The user can correct it by merging or dividing regions, units are region of the initial

oversegmentation. The system does not ask the user for all the segmentations but only for a sample of it. Decisions of validation/correction/rejection are reinjected into the system to guide and improve the segmentation of the other videos. The VO mining is based on description of the VOs, but also needs a feedback from the user. Here, the feedback consists in the presentation of sample results to the user. For example, in classification task, the systems shows samples from some clusters to the user, the user can validate them, correct them or even reject them. To correct them, the user has to move a VO out of a cluster or put a VO into a cluster. Thus, the user guides the processing and at next iteration, the clustering is improved using validation/correction/rejection of the user.

## V. Conclusion

Current VMS deal with videos at the object or region scales but perform mining with shots or videos as elements. In this paper, we have introduced a new taxonomy to characterize VMS and use it to review current object-related VMS showing its relevance to efficiently compare different VMS. We have also asserted that objects should play an even more central role and justify our proposal by presenting the benefits which may be offered by such video object mining systems. We have discussed the impact of choosing objects as elements on the other characteristics defined in our taxonomy. The importance of the VO segmentation step has been highlighted and a way of introducing semantics into VOMS has been proposed. Finally, we have proposed a framework to deal with video object mining (VOMF) and introduced our current work. VOMS offers new prospects and we strongly believe that higher quality and relevance in video mining can be achieved using real (i.e., semantic) objects present in the video.

Future work includes the use of VOMF to design VMS for different objectives. In this perspective, we currently work on the problem of VO clustering. The goal of this research project is to obtain clusters of similar VOs from a video repository. We face some problems when adapting theoretical VOMF to our purpose. Extracting real objects and involving the user in this process is not trivial. In the same way, guiding video mining by user feedback is a complex task. These two issues need more research efforts and experiments.

## Acknowledgements

This work has been supported by Ready Business System, Entzheim, France and the French National Association for Research and Technology (ANRT). We particulary thank Christian Dhinaut from RBS for his support.

## References

[1] YouTube, "http://www.youtube.com/."

[2] I. Koprinska and S. Carrato, "Temporal video segmentation: A survey," *Signal Processing: Image Communication*, vol. 16, no. 5, pp. 477–500, 2001.

[3] A. Rosenfeld, D. Doermann, and D. DeMenthon, Eds., *Video Mining*. Springer, 2002.

[4] K. Cios, W. Pedrycz, R. Swiniarski, and L. Kurgan, *Data Mining A Knowledge Discovery Approach*. Springer, 2007.

[5] F. Idris and S. Panchanathan, "Review of Image and Video Indexing Techniques," *Journal of Visual Communication and Image Representation*, vol. 8, no. 2, pp. 146–166, 1997.

[6] R. Brunelli, O. Mich, and C. Modena, "A survey on automatic indexing of video data," *Journal of Visual Communication and Representation*, vol. 10, no. 2, pp. 78–112, 1999.

[7] D. Brezeale and D. Cook, "Automatic video classification: A survey of the literature," *IEEE Transactions on Systems, Man and Cybernetics-part C: Applications and Reviews*, vol. 38, no. 3, pp. 416–430, 2008.

[8] A. Money and H. Agius, "Video summarisation: A conceptual framework and survey of the state of the art," *Journal of Visual Communication and Image Representation*, vol. 19, no. 2, pp. 121–143, 2008.

[9] W. Ren, S. Singh, M. Singh, and Y. Zhu, "State-of-the-art on spatio-temporal information based video retrieval," *Pattern Recognition*, vol. 42, no. 2, pp. 267–282, 2009.

[10] S. Lefèvre, J. Holler, and N. Vincent, "A review of real-time segmentation of uncompressed video sequences for content-based search and retrieval," *Real-Time Imaging*, vol. 9, no. 1, pp. 73–98, 2003.

[11] A. Anjulan and N. Canagarajah, "Object based video retrieval with local region tracking," *Signal Processing: Image Communication*, vol. 22, no. 7-8, pp. 607–621, 2007.

[12] ——, "A novel video mining system." in *14th IEEE International Conference on Image Processing*. IEEE, 2007, pp. 185–188.

[13] S. de Avila, A. da Luz, and A. de Araujo, "Vsumm: A simple and efficient approach for automatic video summarization," in *15th International Conference on Systems, Signals and Image Processing*, 2008, pp. 449–452.

[14] A. Basharat, Y. Zhai, and M. Shah, "Content based video matching using spatiotemporal volumes," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 360–377, 2008.

[15] F. Chevalier, J.-P. Domenger, J. Benois-Pineau, and M. Delest, "Retrieval of objects in video by similarity based on graph matching,," *Pattern Recognition Letters*, vol. 28, no. 8, pp. 939–949, 2007.

[16] X. Gao, X. Li, J. Feng, and D. Tao, "Shot-based video retrieval with optical flow tensor and HMMs," *Pattern Recognition Letters*, vol. 30, no. 2, pp. 140–147, 2009.

[17] D. Liu and T. Chen, "Video retrieval based on object discovery," *Computer Vision and Image Understanding*, vol. 113, no. 3, pp. 397–404, 2009.

[18] W. Ren and Y. Zhu, "A video summarization approach based on machine learning," *International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, pp. 450–453, 2008.

[19] J. Sivic and A. Zisserman, "Efficient visual search for objects in videos," *Proceedings of the IEEE*, vol. 96, no. 4, pp. 548–566, 2008.

[20] L. F. Teixeira and L. Corte-Real, "Video object matching across multiple independent views using local descriptors and adaptive learning," *Pattern Recognition Letters*, vol. 30, no. 2, pp. 157–167, 2009.

[21] S. Zhai, B. Luo, J. Tang, and C.-Y. Zhang, "Video abstraction based on relational graphs," in *Proceedings of the Fourth International Conference on Image and Graphics*. IEEE Computer Society, 2007, pp. 827–832.

[22] R. Babu, K. Ramakrishnan, and S. Srinivasan, "Video object segmentation: A compressed domain approach," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 4, pp. 462–474, 2004.

[23] B. Toreyin, A. Cetin, A. Aksay, and M. Akhan, "Moving object detection in wavelet compressed video," *Signal Processing: Image Communication*, vol. 20, no. 3, pp. 255–264, 2005.

[24] C.-C. Hsu, H. Chang, and T.-C. Chang, "Efficient moving object extraction in compressed low-bit-rate video," in *Proceedings of the 2006 International Conference on Intelligent Information Hiding and Multimedia*. Washington, DC, USA: IEEE Computer Society, 2006, pp. 411–414.

[25] The Open Video Project, "http://www.open-video.org/."

[26] I. Ruthven and M. Lalmas, "A survey on the use of relevance feedback for information access systems," *The Knowledge Engineering Review*, vol. 18, no. 2, pp. 95–145, 2003.