UDK 004.89

Mikhnova Olena

A TEMPLATE-BASED APPROACH TO KEY FRAME EXTRACTION FROM VIDEO

The following approaches to key frame extraction have been reviewed: boundary-based, motion-based, visual feature-based, based on clustering, matrix factorization and curve simplification. Methods that belong to each group of the approaches (or several groups simultaneously) have been also analyzed. Drawbacks and benefits of each group and method have been highlighted. As a result, a new technique grounding on mathematical morphology and Ritter's approach has been proposed, which aims for simplification of key frame extraction.

Introduction and Related Work

With the wide spread of video information on the web and in personal archives or enterprise databases, many research work has been done to simplify (from one hand) and improve quality (from the other) of video processing, analysis, recognition, understanding and synthesis. The appearance of such web-services as BBC Motion Gallery, YouTube, Google Video and many others facilitates great attention to the research projects that deal with video data mining. To efficiently extract information from these huge video collections new methods of skimming and summarization have been developed. In this paper, we present current techniques of key frame extraction, analyze benefits and drawbacks of all these techniques, and propose a novel approach.

By key frame extraction we mean procedure of choosing a single frame from a video sequence, which is the main one from the point of some chosen parameters. General steps needed for key frame extraction are shown in fig. 1.1, though they may vary from one method to another. This procedure can be used to reduce the amount of information for viewing. Users get an opportunity to evaluate whether a movie is interesting or not, if a clip contains certain event or not. It is also important for automation of video montage (type of video editing involving overlaying of several frames to obtain continuous movie), summarization (selection of static images from video that altogether represent the whole of its content), content-based video indexing and retrieval, video browsing in archives or databases.

Although this area of research has emerged quite recently — namely when multimedia databases expanded to enormous sizes, a human being would have to browse its all life long — nowadays there exist lots of solutions to key frame extraction. This paper is a brief overview of the main of them.

Key frames were first mentioned in literature in the early 90th, when concepts of selection the first, middle or last frame in a scene appeared [1, 2]. Such concepts belong to shot boundary-based approach. Despite of its simplicity, it does not pay any attention to video content.

A more reasonable approach is based on motion analysis (sometimes combined with visual content-based approach). It assumes optical flow or trajectory computation (in addition with visual features — color, texture, etc.) for each frame and then comparing the results using specially proposed metrics. Such kind of approach is widely used [3-5], though it requires huge computational resources and often leads to unreliable results due to the underlying condition of local maximum or minimum search (or similarity threshold choice) [6, P. 80-81]. The latter is also true for the next given group of techniques.

Visual and motion features are also usually applied in the other approaches, for instance, in key frame extraction techniques implemented via clustering. It is the most wide spread type of techniques researched by many contemporary scientists [6, 7].

Another approach relies on curve simplification where video sequence is presented as a trajectory curve in high-dimensional feature space. Research in the area of curve simplification has been done by S. Lim, D. Thalmann, S. Li, M. Okuda, S. Takahashi, E. Bulut, H. Togawa, K. Matsuda, K. Kondo and many others. In the following section the work of E. Bulut et. al. [8] (one of the representatives mentioned above) is discussed as an example.

And the last type of key frame extraction techniques uses matrix factorization to represent frame features. In terms of this approach, video summary is constructed by using techniques such as singular value decomposition (SVD) [9], low-order discrete cosine term (DCT) [10] or any other.

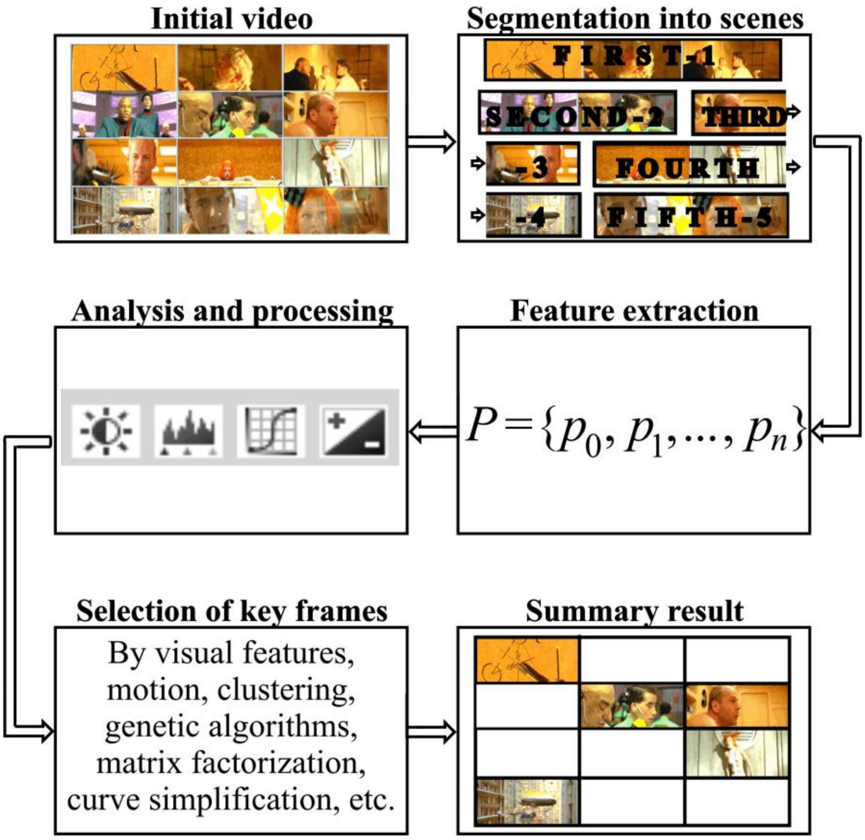

Fig. 1.1. General procedure of key frame extraction

In the next section we observe techniques belonging to each of the above approaches by highlighting pros and cons of their usage. There are also mixed techniques that combine a number of approaches. Of course, there exist techniques that do not belong to any of the above approaches, for instance techniques based on genetic algorithms, deformation analysis

[11], entropy [12], etc. They will not be given below because it is just impossible to cover all of them, so we have chosen only the most prominent ones.

Advantages and Disadvantages of Key Frame Extraction Techniques

In this section we propose to get acquainted with techniques that belong to approaches discussed above. The following table gives an overview of contemporary scientists' contribution to key frame extraction that will be discussed in more details throughout this section.

Table 2.1 Contemporary scientists and their contribution to key frame extraction

Type of Approach	Author (Year of Invention)
Boundary-based	A. Nagasaka, Y. Tanaka (1991)
Motion-based	W. Wolf (1996); T. Liu, H.-J. Zhang, F. Qi (2003)
Based on motion and visual features	Y.S. Avrithis, A.D. Doulamis, N.D. Doulamis, S.D. Kollias (1999)
Motion-based and clustering	X. Zeng, W. Hu, W. Liy, X. Zhang, B. Xu (2008)
Based on clustering	R. Hammoud, R. Mohr (2000); L. Li, X. Zhang, Y. Wang, W. Hu, P. Zhu (2008)
Matrix factorization	Y. Gong, X. Liu (2000)
Curve simplification and clustering	E. Bulut, T. Capin (2007)

The first techniques that appeared were those which belong to boundary-based approach. In 1991 A. Nagasaka et. al. [1] proposed to extract first frame of each scene. These first frames were assigned as the main ones. Despite such methods are fast and easy, the number of selected frames per scene is limited to 1, no matter how meaningful these frames are.

Later, there arise a little bit more sophisticated approaches that take video content into account. In 1996 W. Wolf [3] developed motion-based approach that possesses good identification of gestures which are emphasized by momentary pauses and identification of camera motion. Optical flow is used to find local minimum of motion. His solution is computationally heavy, and it is oriented to salient motion only.

In 1999 Y.S. Avrithis et. al. [5, P. 5, P. 23] proposed to combine motion and visual features. Temporal variation of feature vector trajectory and minimization of a cross-correlation criterion among the frames of each shot are used for key frame extraction. The recursive shortest spanning tree (RSST) algorithm is used for color segmentation of each frame in a given video shot. This method assumes relative computational complexity, and segmentation algorithms are meant to be more accurate.

Alike to the above, in 2003 T. Liu et. al. [4] proposed to combine motion-based temporal segmentation with color-based scene detection. The turning point of motion acceleration and deceleration of each motion pattern is assigned as a key frame. The number of key frames and their location in a given video are determined automatically by motion patterns of video. The proposed approach is threshold free and also fast, but it can only be used for limited number of applications because of the chosen model. Before application of this model to detect key frames, a video sequence must be segmented into scenes.

Key frame extraction with matrix factorization is one more popular group of techniques. Using this approach in 2000, two of its representatives (Y. Gong and X. Liu) [10] proposed to categorize similar frames to common clusters. After singular value decomposition of feature-frame matrix, they obtained refined feature space aided in categorization. This method eliminates duplicates in resulting sequence of frames as each frame is selected from a cluster with similar frames. Instead of scene boundaries, attention is driven to content. Though, initial frame selection for further processing is performed by users, and it is made with fixed time interval which is content independent; frame features are comprised of color histograms only that may lead to irregular frame selection (when frames with nearly the same color distributions and absolutely different content are discarded).

During this period scientists also began to estimate key frames with clustering approaches. Thus, in 2000 R. Hammoud et. al. [6, P.80-81] proposed simple and relatively fast method that groups similar frames within a scene. Temporal variation of color histograms in RGB is modeled via Gaussian mixture density. Bayes information criterion provides automatic selection of appropriate clusters. The only disadvantage is that to decrease computational complexity the number of parameters is reduced. Of course, this influences the results.

Clustering methods were later applied in combination with other methods. For instance, in 2007 E. Bulut et. al. [8] combined curve simplification and clustering approaches. Their method operates in near to real time mode. Curve saliency metric is computed for each point using Gaussian weighted average values (computed at fine (σ) and coarse (2σ) scales) to measure the importance of each frame. This results in numbers of candidate key frames which are later reduced by clustering methods. The only minus is that key frame number is not automatically determined.

In 2008 L. Li et. al. [7] combined motion-based and clustering approaches. They introduced automatic determination of the number of key frames via adaptive k-means. Key frames are detected based on motion features. After features are extracted for each frame, distance matrix is obtained. The number of key frames is then estimated from the distance matrix. Finally, key frames are selected based on k-means clustering. Such features as color, texture, shape can be combined with motion features in their framework. The only drawback to this method is limited number of application.

And the last clustering technique we would like to describe was introduced in 2008 by X. Zeng et. al. [13]. First, similarity matrix is computed, then dominant-set clustering is performed followed by selection of key clusters and, at last, key frame extraction. The uniqueness of this method lies in fact that the restriction of choosing only one frame per cluster has been overcome — the method assumes choosing a number of frames by taking into account scene complexity. Extraction of single frame per scene may not provide appropriate summary because of different length and activity of each scene, as it is usually done in other clustering techniques. In addition, this technique has quite low computational cost comparatively with traditional clustering methods. As for disadvantages, the assumption of frame saliency is made depending only on the fact that the camera focuses more on the most important scenes. When key frames are selected from several long sequences in a cluster, the middle frame in each sequence is assumed to be salient.

Despite the variety of solutions to key frame extraction, it is still very hard to find a good static representation of video sequence that would save the main idea of referred video or just highlight an important event. This is mostly due to different video sources, sizes, quality, cameras used, angles of shooting, and at last video content itself with the goals of referring. None of existing algorithms could deal with all these stuff at once. If one problem is solved, another is left behind. The same drawback for almost all the above approaches is that the number of key frames must be set a priori, and key frames are usually chosen from primarily segmented video — a frame per scene, for instance. That may lead to highly correlated frames as a result or omission of really important frames.

New Application for Template-based G.X. Ritter's Approach

We propose to use G.X. Ritter's approach based on templates and concepts of restriction (reduction to image parts of particular interest), extension (incorporation of smaller images or their parts to a larger ones), domain (frame dimension, i.e. 2-dimensional or 3-dimensional space), and range (variety of values), which has never been applied for key frame extraction purposes yet. The proposed approach generalizes the notion of structuring elements used in mathematical morphology [14-16]. That is why it can be said that such an approach brings great promises in the field because of its simplicity and an ability of being enhanced by (or jointly used with) any other existing methods without significant influence on computational load.

Let a be a frame $a \in F^X$, where F is a set of possible range values of a, and X — the spatial domain of a. In this case a pixel of any frame can be defined as $(x,\ a(x), z)$, where x is the first coordinate or pixel location, $a(x)$ is the pixel value of a at location x, and z — time variable [17, P. 143-144].

In order to incorporate some characteristics (such as color, texture, brightness, intensity or even boundary information of objects) into a frame that would describe it in different ways, we must determine some additional elements. It can be done via templates which are also frames with their pixels presented as frames themselves $t \in (F^X)^Y$, and t is an F^X frame on Y. Thus, parametrized template can be denoted as $t:\ P \to (F^X)^Y$ with the set of regular F-valued templates $\{t(p) \in (F^X)^Y:\ p \in P\}$ [17, P. 145-146]. But when we deal with extraction of main frames we are searching for minimax values of characteristics leaving behind all the rest. For this purpose G.X. Ritter has proposed to use thresholds S. They help in reducing a to a subset of its domain X. Such restriction of a by S can be defined as $a||_S = a \cap (X \times S)$, and from the point of pixel view — as $a||_S = \{(x, a(x)):\ a(x) \in S\}$ [17, P. 147].

Bringing the idea of thresholds is quite common in image processing applications as it aids to lessen a set to the needed subset. In our case this technique can be used for similarity searching and finding repeats in video sequences that match certain threshold. (These similarities or fine repeats can be further removed from video sequence to obtain key frames as the result.) Much like it has been done by X. Yang and Q. Tian [18], S.S. Cheung and

T.P. Ngueyen [19], C. Herley [20] and many others. The only problem with thresholds is that they are usually set for a particular application without generalization, and in fact they cannot be applied to a variety of films with different length and content. For example, a full screen movie with huge changes in scenes and actors should not be treated the same way as small clips of video captured during some technological process in an industrial plant or manufacturing enterprise with the only so called “scene” and “character” inside. That is the point of our further research.

Conclusion

In this paper a number of key frame selection approaches have been observed, some examples of existing techniques (that belong to the described approaches) have been given. Pluses and minuses of each technique have been discussed in details. A new way of searching for salient frames based on mathematical morphology and Ritter’s approach has been proposed, providing benefits of its usage. Though this new technique has not been tested yet, it promises great results due to simplicity and ability of being applied in conjunction with other methods.

Bibliography

1. Nagasaka A. Automatic Video Indexing and Full-Video Search for Object Appearances / A. Nagasaka, Y. Tanaka // Second Working Conference on Visual Database Systems II. — 1991. — P. 113-127.
2. Ueda. H. IMPACT: an interactive natural-motion-picture dedicated multimedia authoring system / H. Ueda, T. Miyatake, S. Yoshizawa // ACM CHI — 1991. — P. 343-350.
3. Wolf W. Key Frame Selection by Motion Analysis. / W. Wolf // IEEE International Conference on Acoustics, Speech and Signal Processing. — 1996. — Vol. 2. — P. 1228-1231.
4. Liu T. A Novel Video Key-Frame-Extraction Algorithm Based on Perceived Motion Energy Model / T. Liu, H.-J. Zhang, F. Qi // IEEE Transactions on circuits and systems for video technology. — 2003. — Vol. 13, No. 10. — P. 1006-1013.
5. Avrithis Y.S. A Stochastic Framework for Optimal Key Frame Extraction from MPEG Video Databases / Y.S. Avrithis, A.D. Doulamis, N.D. Doulamis, S.D. Kollias // Computer Vision and Image Understanding. — 1999. — Vol.5, No 1/2. — P. 3-24.
6. Hammoud R. A probabilistic framework of selecting effective key frames from video browsing and indexing / R. Hammoud, R. Mohr // International Workshop on Real-Time Image Sequence Analysis. — 2000. — P. 79-88.
7. Li L. Nonparametric Motion Feature for Key Frame Extraction in Sports Video / L. Li, X. Zhang, Y. Wang, W. Hu, P. Zhu // Chinese Conference on Pattern Recognition. — 2008. — P. 1-5.
8. Bulut E. Key Frame Extraction from Motion Capture Data by Curve Saliency / E. Bulut, T. Capin // 20th Annual Conference on Computer Animation and Social Agents. — 2007. — P. 63-67.

9. Gong Y. Video summarization using singular value decomposition / Y. Gong, X. Liu. // Computer Vision and Pattern Recognition. — 2000 — P. 174-180.
10. Cooper M. Summarizing video using nonnegative similarity matrix factorization / M. Cooper, J. Foote // IEEE Workshop on Multimedia Signal Processing. — 2002. — P. 25-28.
11. Lee T. Animation Key-Frame Extraction and Simplification Using Deformation Analysis / T. Lee, C. Lin, Y. Wang, T. Chen // IEEE Transactions on Circuits and Systems for Video Technology. — 2008. — Vol. 18, No. 4. — P. 478-486.
12. Pan L. Key Frame Extraction Based on Sub-Shot Segmentation and Entropy Computing / L. Pan, X. Wu, X. Shu // Chinese Conference on Pattern Recognition. — 2009. — P. 1-5.
13. Zeng X. Key-frame Extraction Using Dominant-set Clustering / X. Zeng, W. Hu, W. Liy, X. Zhang, B. Xu // IEEE International Conference on Multimedia and Expo. — 2008. — P. 1285-1288.
14. Serra J. Image Analysis and Mathematical Morphology / J. Serra. — London: Academic Press, 1982. — 610 p.
15. Heijmans H.J.A.M. Mathematical Morphology and its Applications to Image and Signal Processing (Computational Imaging and Vision) / H.J.A.M. Heijmans, J.B.T.M. Roerdink. — Dordrecht: Kluwer Academic Publishers, 1998. — 452 p.
16. Soille P. Morphological image analysis: principles and applications / P. Soille. — 2nd ed. — Berlin: Springer, 2003. — 391 p.
17. Ritter G.X. Image algebra / G.X. Ritter. — Gainesville: University of Florida, 1992. — 392 p.
18. Schonfeld D. Video Search and Mining / D. Schonfeld, C. Shan, D. Tao, L. Wang. — Studies in Computational Intelligence, Vol. 287. — Berlin: Springer, 2010. — 388 p.
19. Cheung S.S. Mining Arbitrary-length Repeated Patterns in Television Broadcast / S.S. Cheung, T.P. Ngueyen // IEEE International Conference on Image Processing. — 2005. — Vol. 3. — P. 181-184.
20. Herley C. ARGOS: Automatically Extracting Repeating Objects From Multimedia Streams / C. Herley // IEEE Transactions on Multimedia. — 2006. — Vol. 8, No 1. — P. 115-129.