UDK 004.932.2

S.V. Mashtalir[1], O.D. Mikhnova[2]

[1] KNURE, Kharkiv, Ukraine, mashtalir_s@kture.kharkov.ua;
[2] KNURE, Kharkiv, Ukraine, elena_mikhnova@ukr.net

STABILIZATION OF KEY FRAME DESCRIPTIONS WITH HIGHER ORDER VORONOI DIAGRAM

Video summary is one of currently developing areas of video mining. Static summary is composed of key frames extracted from video, which fully depict its content. While extracting key frames with the help of Voronoi tessellation comparison, it has been proposed to detail frame content with higher order Voronoi diagrams. This step has lead to simplification of computational procedure compared with increasing the number of initial generator points. Key frames extracted with Voronoi diagrams have been checked for precision and recall and compared with three existing extraction techniques based on optical flow, cluster analysis and curve simplification.

VIDEO SUMMARY, KEY FRAME EXTRACTION, GENERATOR POINT, VORONOI TESSELLATION, HIGHER ORDER VORONOI DIAGRAMS, FINITE SET OF POINTS

Introduction

Last decade is marked by great progress in machine vision based on artificial intelligence techniques. It is mostly due to increase in hardware capacity that made possible to store huge amount of data and process multimedia in a high speed. Just 10 years ago scientists dealt with image processing, on the contrast, contemporary researchers process video, as hardware possibilities are finally opened up. Despite of these favorable conditions, video processing still lacks in high quality uniform methods that might be applicable for a number of subject matters. Aside from variety of application domains, one of the main problems that arises is shooting conditions (for example light, camera move, zoom, etc.). Moreover, video quality also influences on the results. For instance, video shot even at AVCHD lite resolution 1280×720 and high definition video 1920×1080 may show different results for the same processing algorithm. That is why the gap between low level video features (like color and texture) and high level semantics (features that lead to understanding of video content) is so big.

Zhang D., Eakins J.P., Su Z., Hyvonen E., Smith J.R., etc. [1, 2] tried to shorten this gap by incorporating of high level features, but the truth is that the only thing can be done is mid level presentation obtained from interpretation, transformation or filtering of low level data. After transition from image to video retrieval, researchers also continue to analyze the whole video as a sequence of images (consecutive frames), transferring approaches of spatial segmentation. The only thing that changed is added temporal segmentation of video into shots, scenes and subscenes, which group closely related content.

From this point, search for similar content in videos is much more advantageous compared with standalone images. Our research is performed in this context of video recognition. By summarizing video content we propose to extract representative frames that depict information from several similar frames. These similar frames are not obviously consecutive but they must have identical content to be presented by a frame, or this representative frame corresponds to a great change in the scene (outlier).

Key frame extraction may be implemented by different methods, including simple extraction of first/last frame of each scene, motion analysis, clustering analysis, matrix factorization, curve simplification and their modifications observed in [3]. Let's take a look at three most popular groups of methods.

By analyzing optical flow we can only extract frames with significant changes in motion. Moreover, this kind of methods possesses very slow performance especially for high quality data. The results may happen to be very poor for typical motion with homogeneous exposition and scenes. Changes in light may also influence great motion whereas there is no motion at all. Different threshold values should be set for videos with lots of moving objects and poor motion videos. And the main issue for key frame extraction is that optical flow does not reveal significance of motion detected.

Extracting key frames with cluster analyses forces to set the number of clusters a priori as it is impossible to unify cluster assignment procedure for different video types. Similarly to the previous group of methods, cluster analysis badly deals with homogeneous content. In addition, objects are often assigned to wrong clusters. Frames with identical texture and color scheme, but different content, may be treated the same way and not extracted. But unlike the previous group of methods, clustering enables to find centroid which defines frame importance most accurately.

Analyzing key frames extracted by curve simplification, it is notable that many frames extracted are alike. Additional procedures are needed to decline similar frames obtained. Some frames with really different content are not extracted.

For all the mentioned above techniques, the trick is that some users (experts or respondents) do not like to watch wrong frames extracted. They want to get only relevant frames, no matter how small their number will be. These users look for high precision. Others want to

get high recall. They agree to look through wrong frames extracted, they just want to obtain full information and do not miss any true key frame. For this reason it is very hard to estimate the above methods, but a good method assumes that precision gets lower with increase in the amount of key frames extracted.

To find better solution, we propose extracting key frames with the help of Voronoi diagrams and higher order Voronoi diagrams for detailing video content. In the next section we observe formal definitions for Voronoi tessellation and tessellation of higher order from a frame point of view. Section two shows the results of detailing video content with Voronoi diagrams of higher order, and the last section provides numerical estimation for key frame extraction using Voronoi diagrams. Conclusion is given at the end of the paper.

1. Formal Definitions

Initially designed for geodesy, Voronoi diagrams gained large popularity during the past years in computer graphics, especially for 3-D modeling [4, 5]. We propose a novel application for Voronoi diagrams of order-k. The novelty lies in application of Voronoi tessellations as segments for comparison in sequential video to find frames with significant content and extract key frames from video to obtain short static summary. Order-k Voronoi diagrams are to be used for detailing of video frame content.

A simple Voronoi diagram or an order-1 Voronoi diagram V is built on generator points $\{p_1, p_2, ..., p_n\}$ set a priori. The diagram corresponds to decomposition of a set (an image, in our case) into Voronoi tessellations $\{v(p_1), v(p_2), \&, v(p_n)\}$ according to the following rule:

$$v(p_i) = \{z \in \mathbb{R}^2 : d(z, p_i) \le d(z, p_j)\ \forall i \ne j\} \qquad (1)$$

where $d(\circ,\circ)$ – planar Euclidean metric, Voronoi tessellation $v(p_i)$, corresponding to the generator point p_i, includes all the points z, distance to p_i from which is less than distance to the other generator points p_i with index j different from index i [6, 7].

Voronoi diagrams of higher order have been studied by many scientists since 1970. Among the most prominent are the works of Miles, Shamos, Aurenhammer, Agarwal etc. In context of Voronoi diagrams, the term "order" means the amount of generator points that form Voronoi tessellation, and "higher order" or "order-k" assumes that there are more than one generator point (in contrast to simple Voronoi diagrams where one point form a single Voronoi tessellation). Here, "higher order" does not have any relation to dimensionality of space [8], as all the video frames (images) lie in XY plane.

Voronoi diagram of order k, $V^{(k)}$, that is built on n generator points in 2 dimensional space, is a division of a plane into convex polygons, such that points z of each Voronoi tessellation $v_i^{(k)}$ have the same number of nearest generator points p_i, equal to k. The previously mentioned simple Voronoi diagram is a particular case of order-k Voronoi diagram with $k=1$.

To provide a formal definition for Voronoi tessellation of order k, assume that $\{p_1, p_2, ..., p_n\}$ is a set of generator points, and $\{\{p_{1,1}, ..., p_{1,k}\}, ..., \{p_{l,1}, ..., p_{l,k}\}\}$ is a set of subsets with k nearest generator points, then a convex order-k Voronoi polygon $v_i^{(k)}$, formed by generator points $\{p_{i,1}, \ldots, p_{i,k}\}$, can be written the following way:

$$v_i^{(k)} = \{z \in \mathbb{R}^2 : \max\{d(z, p_{i,h}), p_{i,h} \in v_i^{(k)}\} \le \le \min\{d(z, p_{i,j}), p_{i,j} \in V^{(k)} \setminus v_i^{(k)}\}\} . \qquad (2)$$

In other words, the distance between the farthest point of one Voronoi tessellation to its corresponding generator points is closer or equals to the distance to any nearest generator point of another tessellation. Arbitrary Voronoi tessellation of order k may contain from 0 to k generator points, i.e. in a Voronoi tessellation of order k there may be no generator points at all [8, 9].

To compare consecutive video frames presented by Voronoi diagrams, we offer using partition metric $\rho(V', V'')$ introduced in [10]. It provides understanding of how different Voronoi segments in consecutive frames $B'(z)$ and $B''(z)$ (with generator points $\{p'_1, p'_2, ..., p'_n\}$ and $\{p''_1, p''_2, ..., p''_m\}$ respectively) are.

$$\rho^*(B'(z), B''(z)) \approx \sum_{i=1}^{n}\sum_{j=1}^{m} card(v(p'_i)\Delta v(p''_j)) \times \times card(v(p'_i) \cap v(p''_j)) = \rho(V', V''), \qquad (3)$$

where

$$v(p'_i)\Delta v(p''_j) = (v(p''_j) \setminus v(p'_i)) \cup (v(p'_i) \setminus v(p''_j)) .$$

At the first stage of key frame extraction procedure we extract key frames $B_r^*(z) \in S_l(i,j)$ (without limitation in number) within each shot $S_l(i,j) = [B_i(z), B_j(z)]$, $l = 1,2,...$, $i, j \in L_l$, $\sum L_l = Q$ ($q = 1, 2, ..., Q$ is a discrete time of video lasting) according to the rule:

$$r = \arg\min_{r \in L_l} (\sum_{t \in L_l, r \ne t} \rho(B_r(z), B_t(z)) . \qquad (4)$$

At the next stage, key frames from each shot are compared with each other to eliminate repeats or identical key frames.

Shots are detected by adaptive multidimensional time series model described in [11, 12]. Generator points are proposed to be set automatically with one of existing methods mentioned in [13]. During our experiments we have used Harris corner detector to obtain generator points. While performing key frame extraction procedure, area, location and shape have been used as spatial features. Also traditional colour and texture features have been used.

2. Experiments on Detailing Video Content

The increase of order k influences growth in number of Voronoi tessellations, i.e. detailing enhances. Experiments proved that at some step of detailing, it

becomes weaker under gradual increase of order k (see fig. 1). Moreover, detailing starts to fade not when the value of order k reaches its maximum (it reaches maximum when $k = n-1$), but much earlier.

Under $k = n-1$, i.e. when Voronoi diagram order is the highest possible and equals to the total number of generator points minus one, the diagram is called furthest-site Voronoi diagram [9]. It is important to note that the threshold value for detailing reduction is different for different number of generator points.

Fig. 1 shows a frame from Trecvid video sampling collection with order-k Voronoi diagrams for 15 generator points. The order is indicated at the top left of each diagram.

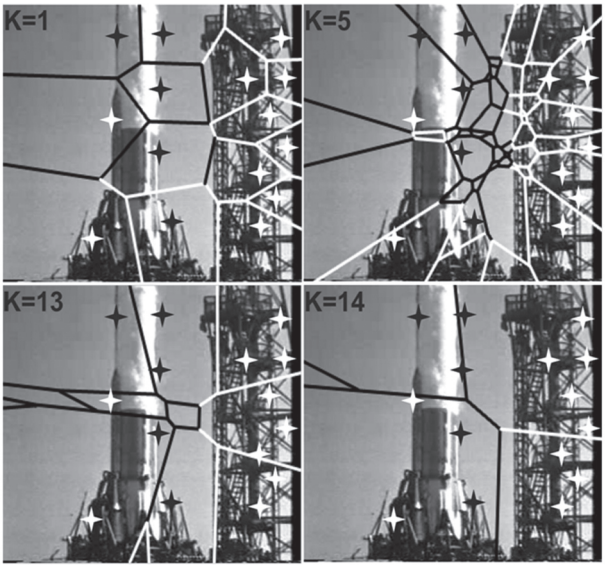

Fig.1. Order-k Voronoi diagrams for 15 generator points

Fig. 2 shows averaged graphic that illustrates the dependence of Voronoi tessellation amount on the order of a diagram with 15 generator points. Though the amount of Voronoi tessellations depends on location of generator points in a plane, experiments held on Trecvid test collection proved that the amount of Voronoi tessellations varies in the interval of 20-30 % for diagrams of the same order with the same number of generator points.

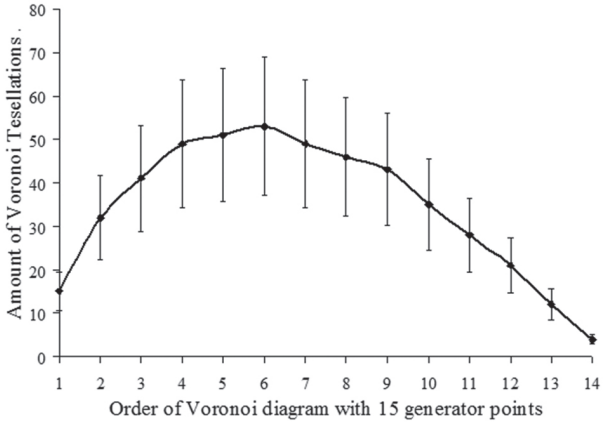

Fig. 2. The dependence of Voronoi tessellation amount on the order of a diagram with 15 generator points

It is important to note that the decrease in detailing usually happens much earlier than a threshold is reached, but, when it is reached, the number of points becomes less in comparison with not only the previous order but also with diagrams of any order including the first order (see fig. 2). In any case, detailing changes under parabola. First, it increases, and then fades smoothly reaching its threshold value near the highest order. Table 1 shows threshold values for 5-35 generator points with 5 point interval. Though, it has been experimentally proved irrational to have more than 20 generator points to construct Voronoi diagrams for further key frame extraction.

Table 1

Threshold values for Voronoi tessellation detailing reduction

Amount of generator points	Order k under which threshold value for detailing reduction is reached
5	4
10	9
15	13
20	18
25	22
30	28
35	32

Thus, Voronoi diagrams of order k bring an excellent tool for detailing of images when they are to be compared with each other. Another variant is to select more generator points initially, but this influences much more computational complexity compared with construction of order-k diagrams. This fact is easily explainable, as generator points are refined and Voronoi diagrams are rebuilt before comparison. The refinement procedure and building of simple order-1 Voronoi diagrams may not converge for a large number of points (sometimes just because of lack in memory resources). Moreover, it is not rational to have so many tessellations for all kind of images and to incorporate detailing in a usual procedure to be performed over time. The more points, the higher order, the greater computational complexity is.

3. Estimation of Key Frame Extraction Procedure

Estimation of results obtained after key frame extraction can be done under subjective opinion of respondents only, which gives an overview of their satisfaction level from video summary they have seen. Such kind of polling results is traditionally studied by the measures of precision and recall. Denote precision by P ($P \in [0;1]$) and recall by R ($R \in [0;1]$). Precision is the amount of found key frames that turned out relevant, and recall is the number of found relevant key frames from all the relevant key frames.

$$P = \frac{tp}{tp + fp};\ R = \frac{tp}{tp + fn}, \qquad (5)$$

where tp, fp, tn and fn can be clearly understood from the following table proposed by [14] while describing textual information search:

Table 2

Notations used for relevant and irrelevant key frames while calculating precision and recall

	Relevant (key frame)	Irrelevant (not a key frame)
Found (*p*)	Extracted key frame (true positive, *tp*)	Not a key frame extracted (fault positive, *fp*)
Not found (*n*)	Not extracted key frame (fault negative, *fn*)	— (true negative, *tn*)

The advantage in usage of two measures simultaneously is obvious. As mentioned above, sometimes it is more important to get small number of relevant frames only, but sometimes it is better to have a huge amount of even fault detected key frames. Recall does not get lower when the number of key frames detected arises.

In total, the task is to reach optimal balance for recall with satisfactory level of fault positive key frames. A measure that finds such a balance is called F-measure. If we put equal weights for precision and recall, we obtain balanced F-measure denoted by F_1 ($F_1 \in [0;1]$) that equals to Dice coefficient [14].

$$F_1 = \frac{2PR}{P+R} \qquad (6)$$

Consequently, Dice coefficient can be used in order to estimate key frame extraction techniques. Fig. 3 illustrates key frames extracted by 4 different methods from Chinese commercial about Mersedes automobiles. For the purpose of estimation, optical flow has been calculated by Horn-Schunck method which analyses changes in motion field energy. It combines optimal balance between good quality and performance according to general rating of optical flow algorithms from Middleburry database [15]. *K*-means has been chosen as clustering method. As for curve simplification method, we have used the same features for it (as for clustering method), but due to difference in procedures the results turned out also different.

Dice coefficient calculated for key frames extracted using Voronoi diagrams from Trecvid sampling collection, from several commercials and self-made high definition test samples (shot at the city centre), equals 0.92 at the average, which is better than Dice coefficient for optical flow method (0.65), and even better than Dice coefficient for clustering (0.78) and curve simplification (0.83) methods. The estimation was performed by 10 independent respondents who did not know the name of key frame extraction method they examined. The highest recall was obtained for optical flow method as too many frames were extracted. The highest precision was obtained for curve simplification method, but some key frames were omitted.

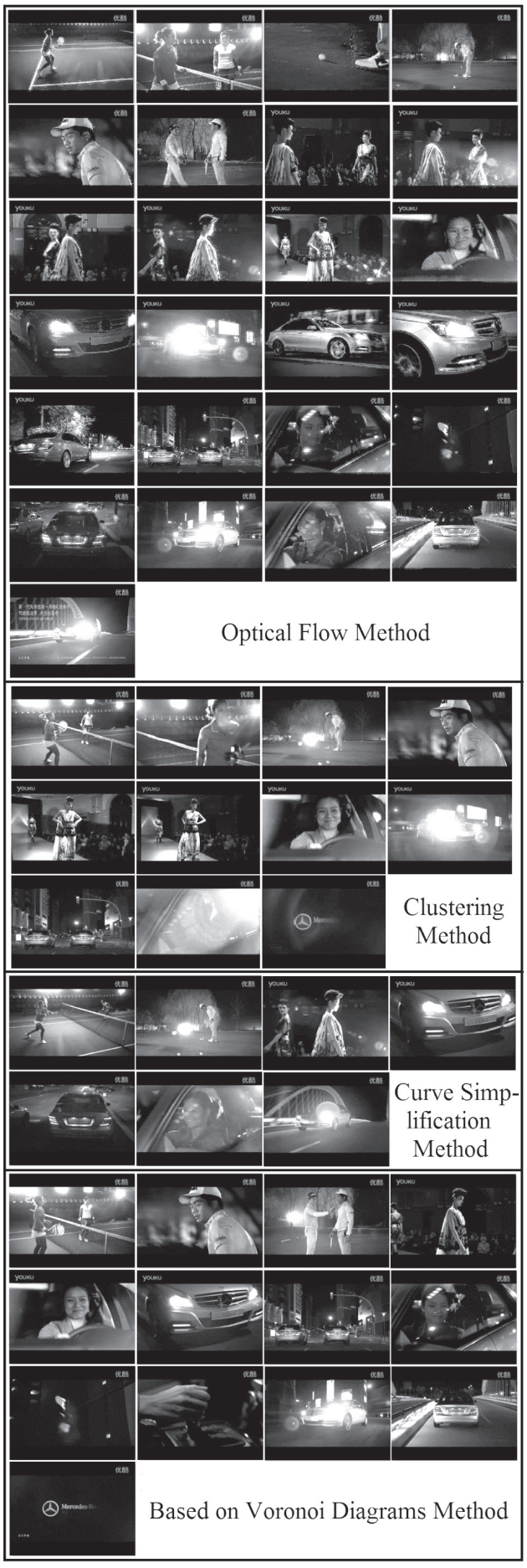

Fig.3. Key frames extracted by 4 different methods

Conclusion

Tasks of video mining are in demand nowadays. Especially it concerns content based retrieval. And it is easily explained, as most of existing methods do not

show man-like processing results. This is true for video summarization. In this paper we propose a novel implementation of higher order Voronoi diagrams, found for geodesy, in static summarization of video. Video frame presentation with Voronoi diagrams and further comparison of them enabled to get machine interpretation of how objects are moved in space and time. Similar content in sequential frames turned out to have identical diagrams. Voronoi diagrams of higher order simplify detailing key frame content compared with increasing the number of initial generator points. Estimation of the proposed key frame extraction procedure shows high precision and recall.

Bibliography: 1. *Zhang D., Liu Y., Hou J.* Digital Image Retrieval Using Intermediate Semantic Features and Multistep Search. In: Digital Image Computing: Techniques and Applications. – 2008. – pp. 513-518. **2.** *Jiang Y.-G., Ngo C.-W., Yang J.* Towards Optimal Bag-of-Features for Object Categorization and Semantic Video Retrieval. In: 6th ACM International Conference on Image and Video Retrieval. – 2007. – pp. 494-501. **3.** *Mikhnova O.* A template-based approach to key frame extraction from video. In: International scientific and technical Internet conf. Computer Graphics and Image Recognition. – Vinnytsia: VNTU. – 2012. – pp. 120-127. **4.** *Ledoux H., Gold C.M.* The 3D Voronoi Diagram: A Tool for the Modelling of Geoscientific Datasets. In: GйoCongrиs. – 2007. – 13 p. **5.** *Carette V. et al.* Modelling and visualisation of fish aggregations using 3D Delaunay triangulation and alpha shapes. In: 8-th International Symposium on GIS and Computer Mapping for Coastal Zone Management. – 2007. – pp. 403-413. **6.** *Du Q., Faber V., Gunzburger M.* Centroidal Voronoi tessellations: Applications and algorithms. In: Society for Industrial and Applied Mathematics Review. – 1999. – Vol. 41, No. 4. – pp. 637-676. **7.** *Hurtado F. et al.* The weighted farthest color Voronoi diagram on trees and graphs. In: Computational Geometry: Theory and Applications. – 2004. – Vol. 27, No. 1. – pp. 13-26. **8.** *Okabe A. et al.* Spatial tessellations: Concepts and applications of Voronoi diagrams. – 2-nd ed. – Chichester: Wiley, 2000. – 671 p. **9.** *Gavrilova M. L.* Generalized Voronoi Diagram: A Geometry-Based Approach to Computational Intelligence. In: Studies in Computational Intelligence. – Berlin: Springer. – 2008. – Vol. 158. – 304 p. **10.** *Mashtalir V. et al.* A novel metric on partitions for image segmentation. In: IEEE International Conference on Video and Signal Based Surveillance. – 2006 – 6 p. **11.** *Bodyanskiy Y. et al.* Adaptive Video Segmentation via Non-stationary Multidimensional Time Series Analysis. In: International Conference on Applied and Theoretical Information Systems Research. – 2012. – 14 p. **12.** *Bodyanskiy Y. et al.* On-line video segmentation using methods of fault detection in multidimensional time sequences. In: International Journal of Electronic Commerce Studies. – 2012. – Vol. 3, No. 1. – pp. 1-20. **13.** *Sebe N., Lew M.S.* Comparing salient point detectors. In: Pattern Recognition Letters. – 2003. – Vol. 24, No. 1-3. – pp. 89–96. **14.** *Manning C.D., Raghavan P., Schutze H.* Introduction to Information Retrieval. – Cambridge: Cambridge University Press, 2008. – 496 p. **15.** *Baker S. at al.* A Database and Evaluation Methodology for Optical Flow. In: International Journal of Computer Vision. – 2011. – Vol. 92, No. 1. – P. 1–31.

Поступила до редколегії 04.12.2012

УДК 004.932.2

Стабилизация описаний ключевых кадров с помощью диаграмм Вороного более высоких порядков / С.В. Машталир, Е.Д. Михнова // Бионика интеллекта: науч.-техн. журнал. – 2013. – № 1 (80). – С. 68-72.

В статье рассматривается актуальное направление распознавания видео с учетом содержимого. Реферирование видео путем извлечения значимый статических изображений, отражающих суть всего материала, является темой исследований. Авторы предприняли попытку реализовать процедуру поиска ключевых кадров с помощью диаграмм Вороного. Диаграммы Вороного более высоких порядков предлагается использовать при детализации содержимого видеокадров.

Ил. 3. Библиогр.: 15 назв.

УДК 004.932.2

Стабілізація описів ключових кадрів за допомогою діаграм Вороного більш високих порядків / С.В. Машталір, Е.Д. Міхнова // Біоніка інтелекту: наук.-техн. журнал. – 2013. – № 1 (80). – С. 68-72.

У статті розглядається актуальний напрямок розпізнавання відео з урахуванням вмісту. Реферування відео шляхом вилучення значимий статичних зображень, що відображають суть всього матеріалу, є темою досліджень. Автори зробили спробу реалізувати процедуру пошуку ключових кадрів за допомогою діаграм Вороного. Діаграми Вороного більш високих порядків пропонується використовувати при деталізації вмісту відеокадрів.

Іл. 3. Бібліогр.: 15 найм.