**UDK 004.932.2**

**S. Mashtalir, O. Mikhnova**
Kharkiv National University of Radio Electronics,
Informatics department

# KEY FRAME RECOGNITION USING VORONOI TESSELLATIONS



**Вилучення ключових кадрів є формою скорочення відеоматеріалу. Цю задачу пропонується вирішувати за допомогою діаграм Вороного, які будуються за опорними точками. Вибирати ці точки пропонується за методом Харріса. Для того, щоб покращити розміщення точок, використовувалась кластеризація методом k-середніх. У результаті опорні точки значно більше відповідають контенту відео, що допомагає, своєю чергою, вилучати лише значущі кадри.**

**Ключові слова: опорна точка, ключовий кадр, скорочення відео, діаграма Вороного, кластеризація k-середніх, метод Харріса**

**Key frame extraction is a form of video summarization. It is proposed to be performed with Voronoi diagrams which are constructed on salient points. Salient point selection is assumed to be done via Harris method. To perfect point location, k-means clustering is used. As a result, salient points much better correspond to video content, which helps to extract meaningful frames.**

**Key words: salient point, key frame, video summarization, Voronoi diagram, k-means clustering, Harris method.**

## Introduction

An explosion in multimedia utilization nowadays has caused the development of generally applicable methods and algorithms for manipulating unstructured or poorly structured data [1]. Video, as part of multimedia, gives the most pertinent and crucially important information. In this respect, video analysis, representation, browsing and retrieval, which are fundamental techniques for accessing video content, yield principal difficulties. Examples of applications include digital video libraries and archives, video-on-demand, interactive TV, communication tools for a distance learning system, video summarization, abstaction and skimming [2]. Moreover, content based video retrieval (CBVR) with query ‘ad exemplum’ can be stated as an integrated problem which covers practically all special tasks of video parsing, processing, archiving, recognition etc. It should be emphasized that, firstly, the major bottleneck of CBVR systems is a large semantic gap between the low-level frame or video features and high-level semantic concepts, secondly, the perception subjectivity problem also challenges such systems, thirdly, de facto it is impossible to process any video as a whole because the amount of data and information properties are enormous [3, 4].

As processing of any video as a whole is impossible, it is extremely preferable to decompose video stream into certain ‘building blocks’. As a rule, video sequences are brought into correlation with chains: video stream – video scenes – video groups – shots – key frames – frames [5]. Major generally accepted blocks of video processing are shots that present consecutive homogeneous sequence of frames recorded from a single uninterrupted camera commonly with fixed position and orientation. Reliability and precision of video parsing process (via time series segmentation of features) are the key to efficient and effective solutions [6], but lap dissolve, fade, side curtain wipe etc. put significant obstacles in feasible shot boundary detection. Along with

appreciable simplification of computational models, shot interpretation in form of key frame(s), i.e. the frames that represent the salient visual content of a shot, provide additional possibilities for video summarization and abstraction. Although considerable progress has been made in key frame production, description and processing, a lot of problems concerning key frame based visual content remain far from being solved.

The paper is organized as follows. Section one discusses problem statement for key frame generation. Section two provides information about salient points used as generator points for Voronoi diagrams. The next section is devoted to refinement of generator points implemented by *k*-means clustering. Validity of obtained clusters has been discussed in the last section. Conclusion is given at the end of the paper.

## Problem statement

Recently video mining has become extremely popular. As a result, there exist tremendous amount of clips (e.g. in the web space) which should be stored and indexed somehow for quick and efficient search. One of such mining tasks is video summarization, namely key frame extraction which is the point our research. The aim is to shorten the material and give a five second glance on what was going on for an hour or more. On the other hand, key frame extraction allows applying existing tools of content based image retrieval for CBVR systems. It is widely accepted that there are two paradigms of key frame detection [7]. The first one consists in shot boundary detection which is interpreted as the key frame. The second one lies in summarization of information in each shot by selecting the most representative frame. It should be emphasized that each frame (and key frame also) is first mapped to a point in a certain feature space, where the features can be categorized into shape, color, texture, etc. Then similarity matching (explicitly or not) represents a basis for video stream parsing and content based retrieval. Consider formal aspects of these problems.

Let $D=[a,b]\times[c,d]$, $a,b,c,d=const$ be a field of view. Denote by $B_k(z), z=(x,y)\in D$ the $k$-th frame from video sequence $\Phi$ (here and subsequently $k=1,2,...,K$ is a discrete time). If $1\le i<j\le K$ and $B_i(z), B_j(z)\in\Phi$ then we shall use notation $S_l(i,j)=[B_i(z), B_j(z)]$, $l=1,2,...$, $i,j\in L_l$, $\sum L_l=K$ for a shot, i.e. ordered in time finite set of sequential frames obtained by temporal segmentation, representing a partition of a video stream into a set of meaningful and manageable segments s.t.

$$\forall l\ S_l(i,j)\neq\varnothing,\ \Phi=\bigcup_{l\in L}S_l(i,j),\ \forall l',l''\ S_{l'}(i,j)\cap S_{l''}(i,j)=\varnothing.$$

For a fixed $l$, define a key frame as an image $B_r^*(z)\in S_l(i,j)$ with property $r=\arg\min_{r\in L_l}(\sum_{t\in L_l,\,r\neq t}\rho(B_r(z),B_t(z))$ where $\rho(\circ,\circ)$ is a measure of dissimilarity (it is desirable to be a metric). Finally, after renumbering we obtain the set $\{B_l^*(z)\}$ of key frames for video stream $\Phi$.

The accuracy and reliability of scene change detection and key frame identification are very significant requirements because general performance of video analysis strongly depends on this stage. Because digital video libraries and archives of immense size are becoming available and due to rich content of video data, feature space required for key frame description has to possess a variety of properties and limitations. One of acceptable approaches lies in concept of image salient points. The points, at which image intensities have visual information carrying geometry, are usually considered salient points. Such points have to satisfy several properties, the main of which are distinctiveness and invariance, i.e. a point should be distinguishable from neighbors and its position should be stable to potential distortions. Förstner operator, Harris point detector, scale invariant feature transform (SIFT) descriptor and many others [8, 9] are typical tools for image salient point detection. However, handled points do not always provide enough feature correlation with image content, so, to perfect salient point detection, regional properties should be considered. Taking into account existing key frame extraction techniques with their advantages and drawbacks [7] note that all the extraction procedures assume frame comparison

to obtain interdependences between them for further decision making. This matching is proposed to be done by comparing of Voronoi diagrams which provide generation of frame partition [10, 11]. These partitions present frames in a rough outline but define descriptions more exactly than salient points only. As a result we have found reasonable compromise between computational complexity and features validity. Matching of Voronoi diagrams may be defined as follows.

Let $\{p_1, p_2, ..., p_n\}$ be a finite set of salient points that we shall use as generator points for Voronoi tessellations. Voronoi diagram is a field of view partition $V = \{v(p_1) \bigcap D, v(p_2) \bigcap D, \ldots, v(p_n) \bigcap D\}$ s.t. $v(p_i) = \{z \in \mathbb{R}^2 : d(z, p_i) \le d(z, p_j)\ \forall i \ne j\}$ where $d(\circ, \circ)$ denotes planar Euclidean metric. Further, consider two frames $B'(z), B''(z)$ with salient points $\{p'_1, p'_2, ..., p'_n\}$ and $\{p''_1, p''_2, ..., p''_m\}$ respectively, then image dissimilarity can be approximately represented by partition metric $\rho(V', V'')$ [12]

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

To obtain Voronoi diagrams, salient points must be set a priori. Currently, there are many techniques for salient point selection [8, 9], that is why we propose to use one of the existing techniques (for instance, the one proposed in [13] or any other corner detector or technique based on wavelets). In the next section Harris method is briefly described. This salient point detector has been chosen as one of the most frequently used in boundary based approach due to its good performance and relative simplicity. Though, as all the rest of the methods, it cannot assume all the image variations. For some kind of images it works better than for others. Actually, this problem arises in any image processing technique.

To overcome or at least partially smooth this problem, salient points (obtained after Harris algorithm implementation) are to be refined with *k*-means clustering algorithm which takes into account such properties as color, texture and relative location of regions in analyzed video frames.

**Salient point selection**

As stated above, salient points are selected with Harris corner detector. In fact, this method does not search for corners, it reveals great changes of intensity in a local Gaussian window $W$. Harris method is based on autocorrelation function that searches for points with great changes of intensity in a frame (as defined above by function $B(x, y)$) in two directions. When two eigen values of Harris matrix possess high values simultaneously [14, pp. 158-160], the algorithm marks salient point at the centre of local window. Harris matrix $A_W(x, y)$ is a second derivative at a point with mentioned above planar coordinates $(x, y)$:

$$A_W(x, y) = \begin{bmatrix} \sum_{x \in W} \sum_{y \in W} \frac{\partial^2 B(x, y)}{\partial x^2} & \sum_{x \in W} \sum_{y \in W} \frac{\partial B(x, y)}{\partial x} \frac{\partial B(x, y)}{\partial y} \\ \sum_{x \in W} \sum_{y \in W} \frac{\partial B(x, y)}{\partial x} \frac{\partial B(x, y)}{\partial y} & \sum_{x \in W} \sum_{y \in W} \frac{\partial^2 B(x, y)}{\partial y^2} \end{bmatrix}. \qquad (2)$$

Fig.1 illustrates salient points detected by Harris method in two video frames, shot at high definition at the city centre. The number of salient points has been restricted to 50, as computational complexity of Voronoi diagrams increases, while quality does not. Moreover, it will be seen from the next sections that the number of points should be even less to obtain well-separated clusters with always converging clustering procedure.

It is clear from the figure above that salient points should be refined. After Harris method implementation they correspond with local intensity outliers too much rather than real image content. This problem is supposed to be overcome in the next section by incorporating clustering procedure after initial Voronoi diagrams are computed.

*Fig. 1. Salient points detected by Harris method*

**Clustering extension of Voronoi algorithm**

After initial salient points have been selected and Voronoi diagrams have been computed for them, salient points are to be refined for better correspondence with image content. For this purpose *k*-means clustering has been chosen, as point membership to one or another cluster is defined by the same rules as point membership of a Voronoi tessellation generated for means. *K*-means clustering belongs to unsupervised learning methods. In addition, *k*-means clustering models are the most popular and well-developed due to least square method lying at its basis [15, p. 16].

*K*-means clustering assumes two steps: assignment of objects to clusters and rearrangement of clusters. In terms of pixel clustering in video frames the following procedure takes place. Each pixel $z$ (with its set of features) of a frame is assigned to cluster $\mathrm{v}(p_i)$, so that the distance to cluster centre $p_i$ is closer than distance to any other cluster centre with index $j$. At the second step, the obtained clusters are rearranged according to the following rule:

$$p_i = \frac{1}{\operatorname{card} v(p_i)} \sum_{z \in \mathrm{v}(p_i)} z\,. \tag{3}$$

In other words, new centroid is assigned with a mean value within each cluster. The process comes to an end when centroid shift is less then predefined threshold (i.e. clusters remain unchanged) or maximum number of iterations is reached [16, p. 40].

Squared Euclidean distance has been used as a measure of proximity to incorporate more weight for distant objects. Three Ohta color features [15, p. 588] have been used as a feature vector for cluster analysis. These are: intensity, red-blue difference and green excess.

$$\begin{aligned} &\text{intensity} = \frac{r+g+b}{3}; \\ &\text{red-blue difference} = r-b; \\ &\text{green\_excess} = (2g-r-b). \end{aligned} \tag{4}$$

In the above formula $r$ corresponds to red component, $g$ corresponds to green component and $b$ corresponds to blue component in an image with RGB color scheme.

Entropy has been used as a texture feature for cluster analysis.

$$E = -\sum_i \sum_j \varphi(i, j) \log_2(\varphi(i, j)), \tag{5}$$

where $\varphi(i, j)$ is $(i, j)^{\text{th}}$ entry of normalized gray-tone spatial dependence matrix [17, pp. 618-619].

Fig. 2 illustrates perfection of Voronoi diagram by refinement of salient points, where video frame content is assumed. Initial Voronoi diagram built on Harris salient points is shown on the left, and Voronoi diagram with refined by *k*-means clustering salient points (cluster centers) is shown on the right.

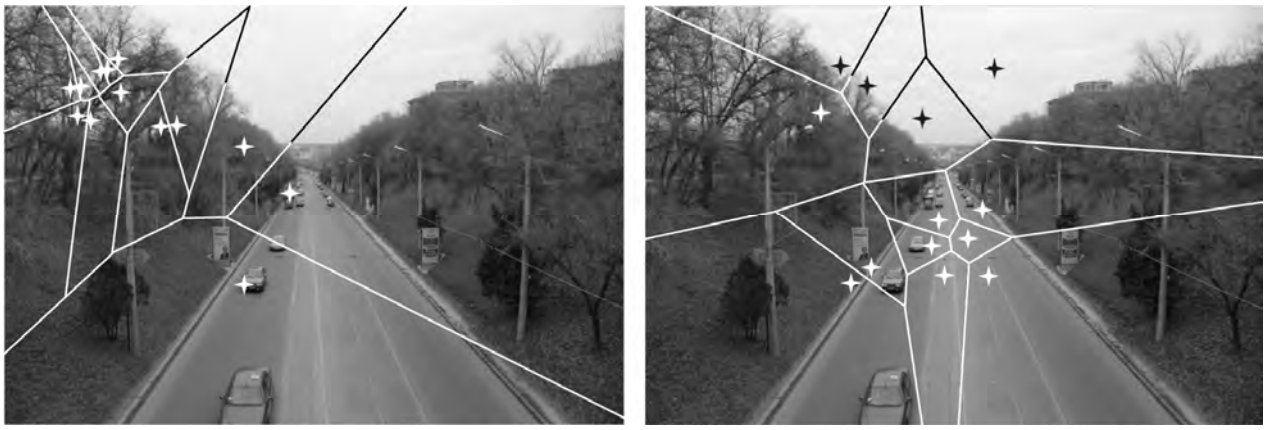

*Fig. 2. Voronoi diagram perfection by refinement of salient points*

The novelty of this approach consists in a fact that the result (shown in fig. 2 on the right) does not depend on initial salient point selection (shown in fig. 2 on the left). In other words, there's no need in optimal point selection. No matter how initial points are located, *k*-means clustering procedure will lead to the same result of arrangement for a particular image. Only number of iterations will differ, but the results will be the same. So, the better Harris method performs, the less iterations will be needed.

**Cluster validity**

To get to know whether clusters are good enough, their validity should be checked. The aim of cluster validity consists in finding some statistical and geometrical properties that characterize cluster accuracy. By statistical indexes researches usually mean matching of obtained clusters to their potential distribution. As a rule, data are distributed under uniform or random distribution. To geometrical indexes belong such properties as volume and density of clusters, distance between them (or their centroids), relation of between/within cluster distances. There are many indexes for between/within cluster distance comparison: F-criterion (1920), Davies and Bouldin index (1979), Bezdek and Pal index (1994), etc. [15, pp. 90-105].

J.C. Bezdek et. al. [15, p. 136] stated that none of existing indexes provide complete estimation assuming a number of clustering algorithms and data structures. To get more accurate results it is recommended using several indexes at once, though it also cannot guarantee total absence of errors. Clustering results are truly correct if only they have been proved by several tests.

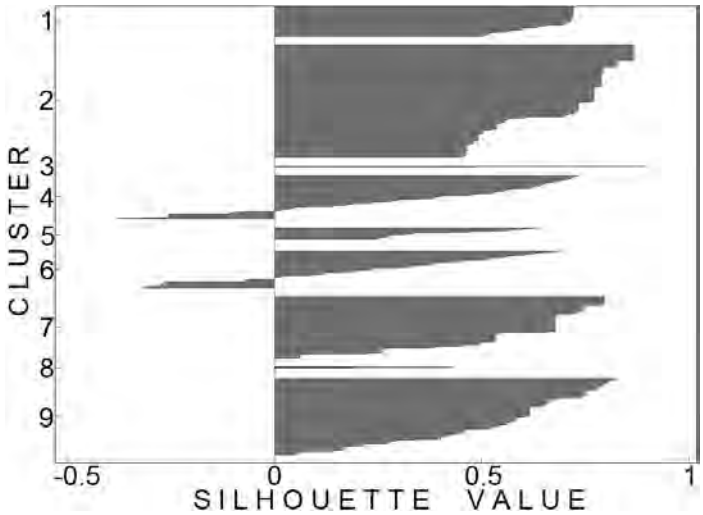


*Fig. 3. Silhouette plot for one of the frames after refinement of salient points with k-means method*

All the mentioned above statistics give numerical presentation of between/within cluster distances. Most of existing statistical software support calculation of average distances within clusters, total sum of distances, Euclidean and squared Euclidean distance between clusters. In our case, silhouette plot will be used for illustration of these parameters. The following graphic shows how close the points of one cluster lie to the points of neighboring clusters. Silhouette value +1 indicates that points of a cluster are most remote to points of neighboring clusters. Silhouette value close to 0 indicates that points of a cluster are too near to the points of neighboring clusters, these points maybe were wrong assigned to their clusters. If silhouette value converges to –1, probably clusters were constructed the wrong way. Fig. 3 shows silhouette plot for one of the frames (from Trecvid collection) after refinement of salient points with *k*-means method.

From the figure above it is clear that only two clusters are well-separated from the others, the rest seven clusters are not that distinct. Most points from the second cluster possess high silhouette values (more than 0.75), that indicates clearness of features at corresponding pixels. However, half points from cluster number 5, 7 and 9 lie at a low level of silhouette value, cluster number 8 totally belongs to interval [0; 0.5), the fourth and sixth clusters have negative values on the graphic. These clusters are not distinct.

Huge amount of negative values and values close to zero are present because of too much number of clusters corresponding to salient points. In fact, these clusters are not characterized by different feature sets, they just lie on a distance from each other at the observed video frame.

The less clusters will be, the more distinct they will be, and the less computations will be needed to build Voronoi diagram. However, if the number of clusters is cut off too much, it may influence the loss of content detailing which as a consequence will lead to wrong key frames extracted. The experiments (hold on Trecvid video samples) have proved that it's not rational to have more than 20 clusters. That is why the number of salient points, detected during Harris procedure, should be limited in the appropriate manner.

## Conclusion

Nowadays, the necessity in video summarization becomes urgent more than ever. To satisfy these needs, a novel technique based on Voronoi diagrams has been proposed for key frame extraction. It takes into account shape changes, texture and color changes, location and area changes. To obtain Voronoi diagrams, salient points must be set a priori. Due to diverse variety of existing algorithms for salient point detection, one of boundary based approaches has been chosen. Being a simple boundary based approach, Harris method can boast its accuracy of results that are very close to those, obtained after wavelet-analyses implementation. But the sad fact is that none of existing image processing methods can deal well with different kind of content. To overcome this problem, clustering extension of Voronoi algorithm has been developed, which leads to invariance of the method used for salient point selection.

After all these work have been done, everything seems to be all right, but great amount of clusters (built on salient points as centroids) leads to misclassification of image pixels. From the other hand, forcible reduction of clusters may influence loss of detailing. Experiments performed on test samples of Trecvid collection have proved that the amount of salient points (and clusters, consequently) should be less than 20. If it would be necessary, it's better to increase detailing by construction of Voronoi diagrams of higher order, rather than increasing the number of clusters, as the number of real world objects (that are interesting for a user or expert) in a frame is usually much less.

*1. Szeliski R. Computer Vision. Algorithms and Applications. – London: Springer, 2011. – 813 p. 2. Goldman D.R. Framework for video annotation, visualization, interaction: Doctoral Thesis … Doctor of Philosophy. – Washington, 2007. – 140 p. 3. 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. 4. 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. 5. Tsoneva T. Automated summarization of movies and TV series on a semantic level: Doctoral Thesis … Doctor of Philosophy. – Eindhoven, 2007. – 126 p. 6. 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. 7. 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. 8. Sebe N., Lew M.S. Comparing salient point detectors. In: Pattern Recognition Letters. – 2003. – Vol. 24, No. 1–3. – pp. 89–96. 9. Tsai Y.-H. Salient points reduction for content-based image retrieval. In: World Academy of Science, Engineering and Technology. – 2009. – Vol. 49. – pp. 656–659. 10. 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. 11. 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. 12. 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. 13. Gonzalez R., Woods R., Eddins S. Digital image processing using MatLab. – Upper Sadle River, NJ: Pearson Prentice Hall, 2004. – 609 p. 14. Sonka M., Hlavac V., Boyle R. Image processing, analysis, and machine vision, International student edition. – 3rd ed. – Toronto: Thomson, 2007. – 850 p. 15. Bezdek J.C. et al. Fuzzy models and algorithms for pattern recognition and image processing. – New York: Springer, 2005. – 776 p. 16. Schonfeld D. et al. (Eds.) Video search and mining. In: Studies in Computational Intelligence. – Berlin: Springer. – 2010. – Vol. 287. – 388 p. 17. Haralick R.M., Shanmugam K., Dinstein I. Textural features for image classification. In: IEEE Transactions on Systems, Man and Cybernetics. – 1973. – Vol. 3, No. 6. – pp. 610–621.*