Computers and Electrical Engineering 39 (2013) 790–796

Contents lists available at SciVerse ScienceDirect

Computers and Electrical Engineering

journal homepage: www.elsevier.com/locate/compeleceng

Sports video summarization based on motion analysis ☆

Engin Mendi [a,*], Hélio B. Clemente [b], Coskun Bayrak [b]

[a] Department of Computer Engineering, KTO Karatay University, Akabe Mah. Cemil Cicek Caddesi, 42020 Karatay, Konya, Turkey
[b] Department of Computer Science, University of Arkansas at Little Rock, 2801 S. University Ave, Little Rock, AR, USA

ARTICLE INFO

Article history:
Available online 20 December 2012

ABSTRACT

Non-annotated video is more common than ever and this fact leads to an emerging field called video summarization. Key frame selection using motion analysis can greatly increase the understanding of the video content by presenting a series of frames summarizing the intended video. In this paper, we present an automatic video summarization technique based on motion analysis. The proposed technique defines motion metrics estimated from two optical flow algorithms, each using two different key frame selection criteria. We conducted a subjective user study to evaluate the performance of the motion metrics. The summarization process is threshold free and experimental results have verified the effectiveness of the method.

1. Introduction

The proliferation of multimedia data such as video, for both entertainment and professional services has increased more than ever in the past few years. This fact has attracted numerous attentions to techniques for automatic video summarization due to its commercial potential. Some of these techniques are becoming more used even in live sports transmissions [1]. A video summary aims to capture the highlights of the video, rapidly providing to the users sense if they will enjoy the movie or not. Furthermore, applying a representation schema such as morphological shape decomposition [2,3], fuzzy feature-based [4] to the video summary containing a set of key frames extracted are very useful for several multimedia applications including content-based video indexing and browsing.

Video shots, sometimes referred to basic scenes [5], are the basic elements of video indexing and difficult to handle most of the time and they are replaced with key frames to be represented. Key frame extraction is fundamental process in video content management. It involves selecting one or multiple frames that will represent the content of the video and used for generating video summaries. Fig. 1 shows hierarchical structure in a video sequence in the extraction of such key frames.

In this paper, we propose a technique for automatic video summarization by motion analysis. Widely used color-based methods do not include motion information in videos. However, motion activity is important especially for sports videos having high action content. We explore the motion features based on optical flow for summarizing sports videos particularly Rugby 7s videos. Motion functions are computed using two different optical flow algorithms and key frame selection criterias. Key frames to produce video summaries are then chosen based on these functions. We tested our method on various Rugby 7s sequences. The rest of this paper is organized as follows: in the next section we give an overview of key frame extraction and video summarization algorithms. In Section 3, we describe our summarization technique. Experiment results are shown in Section 4, and finally, Section 5 concludes the paper.

☆ Reviews processed and approved for publication by Editor-in-Chief Dr. Manu Malek.

* Corresponding author.
E-mail addresses: engin.mendi@karatay.edu.tr (E. Mendi), hbclemente@ualr.edu (H.B. Clemente), cxbayrak@ualr.edu (C. Bayrak).

0045-7906/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.compeleceng.2012.11.020

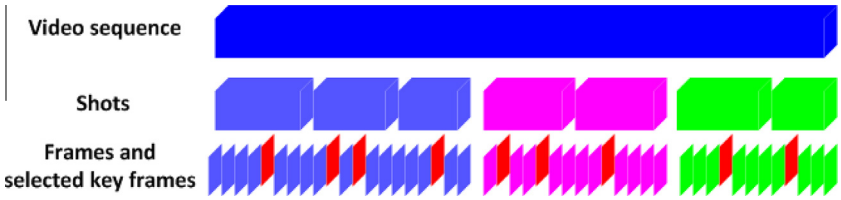

Fig. 1. Hierarchical structure of a video sequence [6].

2. Background

Shots are difficult to handle most of the time and they are replaced with key frames to be represented. Key frame extraction provides a compact pictorial summarization and representation of a video sequence. Video summarization enables navigation and browsing the original video [7,8].

It is worth noting that the choice of key frames to generate a summary is subjective and often application dependent [9]. One of the most straightforward approaches is choosing the first frame of a continuous shot as a key frame [10]. Ardizzone and Cascia [11] selected the key frames based on the length of the shot. If a shot is shorter than one second then only one key frame is selected. Otherwise, for each second one frame is selected as a key frame. An alternative approach is presented by Vermaak et al. [12] based on a frame utility function. Extracting features from video sequence, a utility function is designed that assigns high reward to subsequences of key frames that are maximally distinct and individually carry the most information. Then, for a specified level of detail and endpoints the key frame sequence is obtained that maximizes this utility function by a non-iterative dynamic programming procedure.

Another main approach to key frame extraction is matrix factorization. The frames of video sequences are represented as matrices. Then, by applying a matrix factorization technique to this feature-frame matrix key frames are selected. Gong and Liu [13] proposed a key frame extraction method using singular value decomposition (SVD). Three dimensional histograms in RGB color space with 125 bins are used to create feature-frame matrix. Performing SVD on this matrix, optimal set of key frames is generated based on the visual content metric derived from the SVD properties. Non-negative similarity matrix factorization [14] and sliding-window SVD [15] are other approaches for key frame extraction based on matrix factorization.

Curve simplification methods are also popular in key frame selection problem. A simplified curve is computed that approximates a trajectory curve representing a video sequence in a high dimensional feature space, according to some predefined error criterion. The junctions between simplified curve segments are then chosen as key frames. A recursive multidimensional curve splitting algorithm is proposed in which video curve is represented as a tree structure [16]. Perceptually significant points on the video curve are connected by straight lines and key frames are selected at those significant points as different levels of detail. A major drawback of this approach reported by Li et al. [17] is the difficulty in evaluating the applicability of obtained key frames due to the lack of comprehensive user study to prove. Zhao et al. [18] suggested that a better approximation may be obtained by breaking the entire curve at sharp corners. Then the corners are selected as key frames.

Chang et al. [19] adopted graph theory into key frame selection process, considering a shot as proximity graph and vertices as frames in the graph. It is proposed that key frame extraction problem turned to the optimization problem finding a smallest vertex cover in the constructed proximity graph, that minimizes the total feature distance between the vertices and their neigboring points. Since vertex cover problem is an NP-complete problem which means no fast solution is known, a greedy method is employed whose computational cost is significantly lower.

Recent advances in key frame selection also utilize the neural network models. Narasimha et al. [20] used a neural network to select key frames having high intensity and maximum spatial activity at the center of the frame by extracting MPEG-7 descriptors namely motion intensity and spatial activity descriptors. Li and Doermann [21] used wavelets for training the neural network instead of MPEG descriptors. Cecen [22] applied self organizing map (SOM) on feature vectors created by discrete cosine transform (DCT) coefficients. Key frames are then chosen based on the minimum distance between corresponding neurons and number of winning times.

With respect to the drawbacks or disadvantages associated with each method presented above, the automatic video summarization technique based on motion analysis defines motion metrics estimated from two different optical flow algorithms and using two key frame selection criteria. The advantage of proposed approach is capturing motion information which is crucial for videos containing dynamic content, particularly sports videos. Therefore, our technique can provide more meaningful summary by capturing the high action content through motion analysis. In addition, while most of existing summarization techniques requires thresholding operation, our approach is not threshold-based.

3. Summarization

Shots in the video sequences are segmented using a difference metric based on the color histograms of successive frames [23–25]. Each shot was then processed using the phases described in this section.

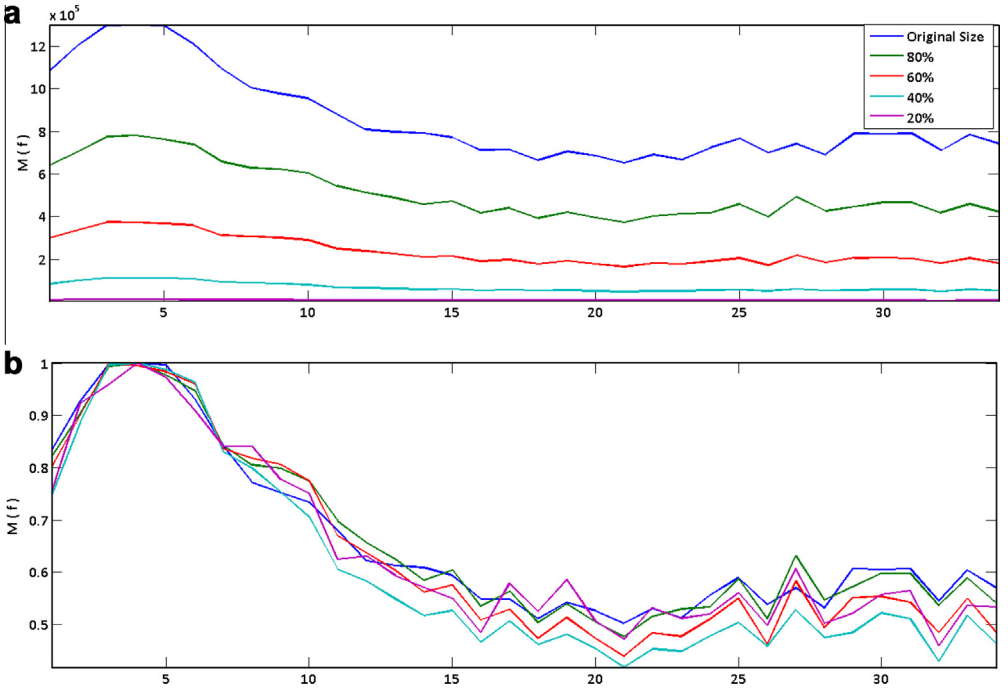

Fig. 2. (a) Motion for different size frames, and (b) normalized motion.

3.1. Pre-processing

Frames are first pre-processed by reducing their sizes to speed up the computation during optical flow estimation. Fig. 2 shows the motion metric calculated from the optical flow with different image sizes. As shown in Fig. 2a, the amplitude varies for the different images sizes, while the shape of the motion function remains same. Since the choice of the key frames will be based on the shape of the motion function and there is no significant change in normalized motion (Fig. 2b), a smaller image size can be used. Images are then converted from the RGB color space to grayscale intensity, removing hue and saturation information while retaining the luminance.

3.2. Motion function

Frames are compared based on optical flow by estimating the brightness patterns between two frames. Major advantages of optical flow are that it is simple and computational cost is low [26]. They also often work effectively for flow estimation due to hand motion and noise [27]. To calculate optical flow, Lucas and Kanade [28] and Horn and Schunck [29] algorithms were used. Lucas–Kanade is classified as a local, while Horn–Schunck is a global method. Lucas–Kanade algorithm is based on the sum of squared error between two images. It assumes that the velocity of all pixels in a region is constant and minimizes:

$$\sum_{p} W^2(p)[\nabla^T I(x_i,t)\cdot v_f + I_t(x_i,t)]^2 \tag{1}$$

where p is a pixel point of a spatial local region, W is the weighted function, $I(x,y,t)$ is image brightness at location (x,y) at time t and $v_f = (h_f, v_f)$ is image velocity. By the derivative of (1) with respect to v_f, the estimated velocity is obtained as:

$$v_f = -A^{-1}r \tag{2}$$

where

$$A = \begin{bmatrix} \sum W^2 I_x^2 & \sum W^2 I_x I_y \\ \sum W^2 I_x I_y & \sum W^2 I_y^2 \end{bmatrix}$$
$$r = \begin{bmatrix} \sum W^2 I_x I_t \\ \sum W^2 I_y I_t \end{bmatrix} \tag{3}$$

and

$$|A| \neq 0 \tag{4}$$

In Horn–Schunck algorithm, the gradient constraint is combined with a global smoothness term to constrain the estimated velocity field. Horn–Schunck indicates such smoothness constraint by optical flow derivative and minimizes:

$$\iint (I_x h_f + I_y v_f + I_t)^2 + \alpha(|h_f|^2 + |v_f|^2)dxdy \tag{5}$$

where α is smoothness coefficient. Minimizing the functional (5), following partial differential equations are obtained:

$$\begin{aligned} \alpha\Delta h_f - I_x(I_x h_f + I_y v_f + I_t) = 0 \\ \alpha\Delta v_f - I_y(I_x h_f + I_y v_f + I_t) = 0 \end{aligned} \tag{6}$$

which suggests following iterative scheme to solve:

$$\begin{aligned} h_f^{k+1} &= \bar{h}_f^k - I_x \frac{I_x \bar{h}_f^k + I_y \bar{v}_f^k + I_t}{\alpha + I_x^2 + I_y^2} \\ v_f^{k+1} &= \bar{v}_f^k - I_y \frac{I_x \bar{h}_f^k + I_y \bar{v}_f^k + I_t}{\alpha + I_x^2 + I_y^2} \end{aligned} \tag{7}$$

where k denotes last calculated result, $k+1$ is next iteration, $\bar{h}_f$ and $\bar{v}_f$ denote average of neighboring values of h_f and v_f, respectively.

Proposed motion metric function is calculated by adding the magnitudes of velocity components from optical flows at each pixel:

$$M(f) = \sum_i \sum_j |h_f(i,j)| + |v_f(i,j)| \tag{8}$$

where $h_f(i,j)$ and $v_f(i,j)$ are the vertical and horizontal velocity fields of frame f at pixel locations i and j. To remove noise, motion function is smoothed using a weighted linear least squares regression and a second degree polynomial model. Fig. 3 shows a motion function with noise and smoothed motion function of a video segment obtained from Lucas–Kanade optical flow.

3.3. Key frame extraction

Key-frame extraction is the essence of video summarization. Extrema points carry important motion information for the high action content. To choose the key frames from the smoothed motion function, we have used two approaches: We first select the key frames from the global extrema (maximums and minimums) in the motion function. Using this approach, we obtain two key frames in every shot. Second, we select the key frames at the local minimum between two maximums. We fix

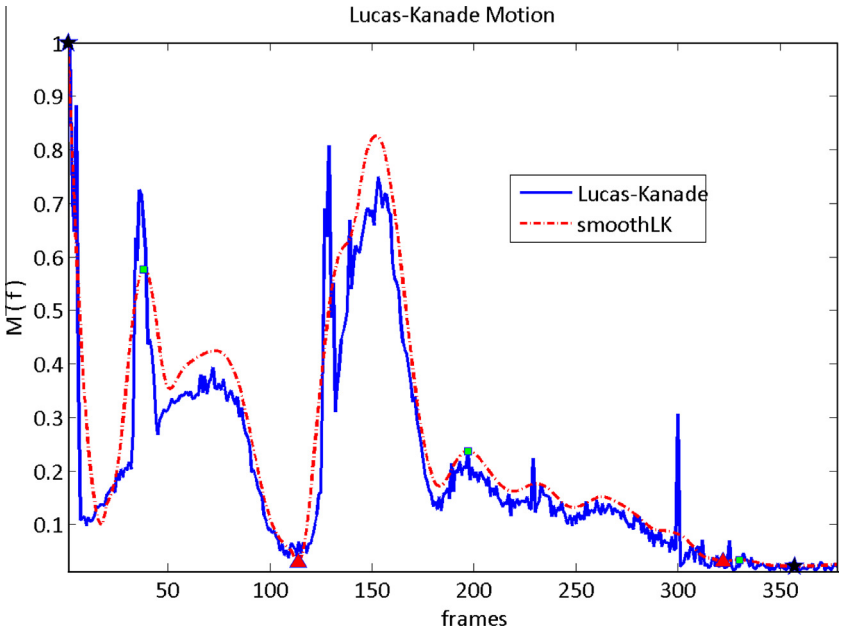

Fig. 3. Noisy and smoothed motion function.

the first local maximum and search the next maximum having a difference of N with the magnitude of the first one. The local minimum is chosen as a key frame. We then repeat this process until finding another local maximum. This approach does not assume a fixed key frame per shot. In Fig. 3, key frames obtained using global extrema are marked with $*$ and local minimum between the two maximums with $N = 0.5$ are marked with Δ.

4. Experimental results

Combining motion metrics derived from two optical flow algorithms with two approaches described in the previous section, we obtain four metrics: (i) Horn–Schunck using global extrema, (ii) Lukas–Kanade using global extrema,

Fig. 4. Video summaries of Rugby 7s video using (a) selecting middle frame; our motion metric derived from (b) Horn–Schunck using global extrema, (c) Lukas–Kanade using global extrema, (d) Horn–Schunck using local minimum between two maximums, and (e) Lukas–Kanade using local minimum between two maximums.

(iii) Horn–Schunck using local minimum between two maximums, and (iv) Lukas–Kanade using local minimum between two maximums. We also compare our metrics with one of the most commonly used method which is selecting the middle frame of a shot as its key frame due to its simplicity. Fig. 4 shows the results of selecting middle frames and applying the proposed motion metrics on a Rugby 7s video.

As there is no objective ground truth or benchmarking result for performance measurement of video summarization algorithms, we have conducted a subjective user study for evaluating the proposed techniques. Truong and Venkatesh [30] claimed subjective user studies are the most useful and realistic form of evaluation in video summarization.

We have developed a web-based interface presenting a video and corresponding summaries to be evaluated (Fig. 5). Twelve testers participated in the evaluation process. They were asked to rate the summaries based on representativeness, coverage and redundancy. For each shot, three satisfactory scores (Good, Reasonable or Bad) are used.

The test data is composed of 46 shots and more than 1 h sports videos, in particular Rugby 7s. 40% of test data is short shot, 60% is long shot. If a shot is shorter than 5 s then it is considered as a short shot, otherwise a long shot. The results of the evaluation for short and long shots are shown in Tables 1 and 2, respectively. The rows of the tables show the percentages of Good, Reasonable and Bad rates given by the testers over total test shots. The columns of the tables show the methods. Method 1 refers to middle frame approach. Methods 2, 3, 4 and 5 refer to our metrics. Method 2 and Method 3 are the metrics derived from Horn–Schunck and Lukas–Kanade using global extrema, respectively. Method 4 and Method 5 are the metrics derived from Horn–Schunck and Lukas–Kanade using local minimum between two maximums, respectively.

We can see from tables that Method 4 and Method 5 perform better for short shots while Method 1, Method 2 and Method 3 are better for long shots. From the results, two conclusions can be drawn. First, for videos containing short shots, generated summaries by choosing the key frames at the local minimum between two maximums are more representative and nonredundant. Second, selecting key frames at the global extrema can better capture the salient visual content within a video sequence containing long shots. Considering aforementioned observations, it is possible to conclude that key frame selection criteria on

Fig. 5. Web interface for subjective evaluation.

Table 1
Evaluation results of short shots.

	Method 1 (%)	Method 2 (%)	Method 3 (%)	Method 4 (%)	Method 5 (%)
Good	13	15	18	41	43
Reasonable	50	42	40	38	33
Bad	37	43	42	21	25

Table 2
Evaluation results of long shots.

	Method 1 (%)	Method 2 (%)	Method 3 (%)	Method 4 (%)	Method 5 (%)
Good	27	21	19	23	23
Reasonable	41	46	47	34	34
Bad	32	33	36	42	43

the motion profile is more important than motion function of the optical flow in constructing coherent summary of a video clip. Furthermore, the decision of key frame selection criteria must be taken based on the duration of the shots.

5. Conclusion

This paper has presented a video summarization technique based on motion analysis. The method uses optical flow computations to identify global extrema and local minimum between two maximums in the motion. We have tested our approach on Rugby 7s videos. Experimental results have shown that our method is effective and generated summaries are representative. In addition, proposed technique is threshold free and resulting key frames highly depend on the perceived motion patterns of the video.

References

[1] Wang J, Parameswaran N. Survey of sports video analysis: research issues and applications. In: Proceedings of the Pan-Sydney area workshop on visual information processing; 2004. p. 87–90.
[2] Vizireanu DN. Generalizations of binary morphological shape decomposition. J Electr Imaging 2007;16:013002.
[3] Vizireanu DN. Morphological shape decomposition interframe interpolation method. J Electr Imaging 2008;17:013007.
[4] Doulamis AD, Doulamis ND, Kollias SD. A fuzzy video content representation for video summarization and content-based retrieval. Signal Process 2000;80(6):1049–67.
[5] Barbu T. Novel automatic video cut detection technique using Gabor filtering. Comput Electr Eng 2009;35(5):712–21.
[6] Mendi E, Bayrak C. Shot boundary detection and key frame extraction using salient region detection and structural similarity. In: 48th ACM southeast conference (ACM-SE '10), Article 66, Oxford, Mississippi, USA, April 15–17, 2010.
[7] Mendi E, Bayrak C. Summarization of MPEG compressed video sequences. Adv Sci Lett 2011;4(11–12):3706–8.
[8] Mendi E. Automated content-based video analysis and management. Ph.D. dissertation, University of Arkansas at Little Rock, Little Rock, AR, USA; 2012.
[9] Porter SV. Video segmentation and indexing using motion estimation. Ph.D. thesis, University of Bristol, February 2004.
[10] Chen H-Y, Wu J-L. A multi-layer video browsing system. IEEE Trans Consum Electron 1995;41(3):842–50.
[11] Ardizzone E, Cascia ML. Video indexing using optical flow field. In: IEEE international conference on image processing; 1996. p. 831–4.
[12] Vermaak J, Perez P, Blake A, Gangnet M. Rapid summarization and browsing of video sequences. British machine vision conference; 2002. p. 424–33.
[13] Gong Y, Liu X. Video summarization using singular value decomposition. In: Proc. of CVPR; 2000. p. 174–80.
[14] Cooper ML, Foote J. Summarizing video using non-negative similarity matrix factorization. In: IEEE workshop on multimedia signal processing; 2002. p. 25–8.
[15] Almageed WA. Online, simultaneous shot boundary detection and key frame extraction for sports videos using rank tracing. ICIP08; 2008. p. 3200–3.
[16] DeMenthon D, Kobla V, Doermann D. Video summarization by curve simplification. In: ACM multimedia; 1998. p. 211–8.
[17] Li Y, Zhang T, Tretter D. An overview of video abstraction techniques. Technical report HPL-2001-191, HP Laboratory technical report; 2001.
[18] Zhao L, Qi W, Li S, Yang S, Zhang HJ. Key-frame extraction and shot retrieval using nearest feature line. In: Proceedings of ACM multimedia workshop; 2000. p. 217–20.
[19] Chang HS, Sull S, Lee SU. Efficient video indexing scheme for content based retrieval. IEEE Trans Circ Syst Video Technol 1999;9(8):1269–79.
[20] Narasimha R, Savakis A, Rao RM, de Queiroz RL. A neural network approach to key frame extraction. SPIE-IS&T Electr Imaging Storage Retrieval Methods Appl Multimedia 2004;5307:439–47.
[21] Li H, Doermann D. Automatic identification of text in digital video key frames. In: ICPR, August 1998.
[22] Cecen S. Histogram based video segmentation and key frame extraction on SOM and DFT. Master's thesis, University of Arkansas at Little Rock; 2009.
[23] Yu J, Srinath MD. An efficient method for scene cut detection. Pattern Recognit Lett 2001;22:1379–91.
[24] Mendi E, Bayrak C. Shot boundary detection and key frame extraction from neurosurgical video sequences. Imaging Sci J 2011.
[25] Mendi E, Bayrak C. A web-based medical video indexing environment. In: 4th IEEE international conference on semantic computing (ICSC '10), Pittsburgh, PA, USA, September 22–24, 2010. p. 172–5.
[26] Hayakawa H, Shibata T. Block-matching-based motion field generation utilizing directional edge displacement. Comput Electr Eng 2000;36(4):617–25.
[27] Lakshman P. Combining deblurring and denoising for handheld HDR imaging in low light conditions. Comput Electr Eng 2012;38(2):434–43.
[28] Lucas BD, Kanade T. An iterative image registration technique with an application to stereo vision. In: Proceedings of imaging understanding workshop; 1981. p. 121–30.
[29] Horn BKP, Schunck BG. Determining optical flow. Artif Intell 1981;17(8):185–203.
[30] Truong BT, Venkatesh S. Video abstraction: a systematic review and classification. ACM Trans Multimedia Comput, Commun, Appl 2007;3(1).

Engin Mendi received his Ph.D. degree in Integrated Computing with Computer Science track from University of Arkansas at Little Rock (UALR), two MS degrees in Applied Science from UALR and in Computational Engineering from Technical University of Munich and BS degree in Civil Engineering from Middle East Technical University. His research interests are in image processing, artificial intelligence, computational neuroscience, computer vision, and data mining.

Hélio B. Clemente received his BS degree in Engineering Sciences and Information Technologies and Electronics from Universidade do Algarve, Portugal. His research interests are in image processing, software engineering, telecommunications, automation, and control.

Coskun Bayrak holds a BS from Slippery Rock University of Pennsylvania, and a MS from Texas Tech University, and PhD from Southern Methodist University in Computer Science. His research is in software engineering, component based development, data mining, and biomedical engineering, modeling and simulation, cellular automata, mobile application development, and health care application development.