

SEMANTIC AUDIOVISUAL ANALYSIS FOR VIDEO SUMMARIZATION

Junyong You, *Member, IEEE*, Miska M Hannuksela, *Member, IEEE*, Moncef Gabbouj, *Senior Member, IEEE*

Abstract: This paper proposes a semantic audiovisual analysis approach for video summarization. The sequence to be analyzed is first segmented into scenes according to audio similarity. Some global clues such as loudness, the ratio of unrelated shots, and the affective relationship between the scenes and the whole sequence are employed to compute the semantic scene importance. The shots in each scene are grouped based on the luminance histograms, and the semantic shot importance is then calculated using selected audio and video features. Subsequently, key frames are extracted according to the semantic frame importance computed based on certain visual features, such as attention region and motion information. This approach is effective to generate a representative video summary whilst avoiding some disadvantages of the traditional video summarization methods. Experimental results demonstrate promising performance of the proposed approach.

Index Terms: Semantic analysis, audio classification, audio segmentation, video summarization

I. INTRODUCTION

As digital video collections are growing rapidly, information redundancy of audiovisual contents is increasing at an even higher pace. Consequently, automatic extraction of the most representative excerpts of the content is becoming of the key techniques for managing and accessing a large video library. A great deal of video summarization algorithms have been proposed, for example, using video segmentation [1][2], methods based on human attention or perception [3][4], an affective model [5] which maps video contents into human emotion, and highlight selection [6]. Generally, video summaries can be classified into two categories: a series of individual key frames with or without audio, and a series of short excerpt clips of the sequence with audio. Most existing summarization algorithms compute one or more time-varying values that represent particular audiovisual features. Then, key frames or clips corresponding to the greatest values of these feature functions are extracted as the summarization. For example, based on some audio and video features, the algorithms in [3] and [4] estimate the attention and perception viewers may pay to video contents, and then the frames or segments which can attract most attention will be taken as video summary. However, these algorithms usually take into account the local audiovisual features, i.e. the features are computed based on only the current frame and a couple of the closest frames while the relationship between the frame/segment and the whole sequence is neglected. Thus, the computed peak points are usually

the local maxima such that the corresponding frames/segments might not be the most important. Another disadvantage is that many duplicate frames even in one scene are contained in the summarization. In this study, we propose a video summarization method that selects key frames based on semantic analysis for scene, shot, and frame in a holistic structure, which can avoid these disadvantages compared to the traditional methods.

Video can be regarded as a hierarchical structure of sequence \rightarrow scene \rightarrow shot \rightarrow frame. A scene represents the action in a single location, while a shot is taken with a single camera. Although many algorithms for scene segmentation using audio and video features have been presented [7], we make use of audio content to segment a sequence into scenes in order to keep the computational complexity reasonably low. For each scene, shots are segmented using the difference of luminance histogram between the adjacent frames. Subsequently, all the shots in one scene are clustered into different groups with different visual contents according to the similarity between the average histograms of each shot. After that, the semantic importance of segmented scenes, shots, and frames is modeled based on certain audiovisual features, such as motion, audio classification, loudness, shot change frequency, and attention region, which were selected based on the principles of human perception and cinematic techniques. It is proposed that the video summarization is generated according to the semantic importance of a scene, shot, or frame. Thus, the proposed approach is able to extract the most representative excerpts whilst excluding unimportant video segments from the summarization. In addition, although we made the experiments based on the uncompressed audiovisual sequences, all the proposed features can be derived from the compressed streams. Hence, the proposed algorithm is applicable in a real time video analysis system.

This paper is organized as follows: section II gives the detailed description of the proposed semantic analysis approach. Section III presents the generation of video summaries. Experimental results are provided and evaluated in section IV. Conclusions are drawn in section V.

II. SEMANTIC IMPORTANCE ANALYSIS OF AUDIOVISUAL CONTENT

A. Audio scene segmentation and classification

We assume that the audio track of a scene is consistent in terms of the signal characteristics, so a

scene can be segmented using some audio features. In this study, the similarity of the audio features is utilized to segment the scenes directly and the classification is achieved by a GMM (Gaussian mixture model).

As an important index in psychoacoustics, loudness can express the ability to catch human attention. Thus, the loudness is computed first for each audio frame which is defined as 2048 continuous samples with 50% overlap. The frames with low loudness (<3dB), defined as silence, are eliminated first. And then, the segmentation and classification are performed on the remaining frames. The audio signal is divided into clips with the constant length (1 second in this work), then the segmentation and classification are performed on the non-silent clips using the following features:

Volume Based [7]: STE (short time energy), VSTD (volume standard deviation), VDR (volume dynamic range);

ZCR Based [7]: ZCR (zero crossing rate), ZSTD (standard deviation of ZCR);

Pitch Based [7]: PC (pitch contour), PSTD (standard deviation of pitch), SPR (smooth pitch ratio), NPR (non-pitch ratio). In this work, we use the average magnitude difference function (AMDF) to detect the pitch.

Frequency Based [7]: BW (bandwidth), FC (frequency centroid).

MFCC Based [8]: MFCC (Mel-frequency cepstral coefficient), delta MFCC, autocorrelation MFCC.

Scene change is detected based on the above features without MFCC based, using the following change index [9]:

$$SCI = \frac{\left\| \left(\sum_{i=N}^{-1} f(i) - \sum_{i=0}^{N-1} f(i) \right) / N \right\|^2}{\sqrt{(c + \text{avg}(f(-N), \dots, f(-1))) \cdot (c + \text{avg}(f(0), \dots, f(N-1)))}} \quad (1)$$

where $f(i)$ is the feature vector of the i -th clip, while $i=0$ representing the current clip, $\|\bullet\|$ is the $L2$ norm, $\text{avg}(\dots)$ is the average of the squared Euclidean distances between each vector and the mean feature vector of the N clips considered, and c is a small constant to prevent division by zero. Here $N=6$ is a satisfactory selection for scene change detection.

The audio class is predefined as one of five genres: speech, music, noise, silence, and others, where silence is detected by loudness, and the classification of the other four genres is achieved based on the above features by GMM. In fact, it is not necessary to use all the above features, in this study, we make use of STE, ZCR, PC, and MFCC based features to compute the GMM parameters using the training audio data.

B. Shot clustering and semantic scene importance

After scene segmentation, the semantic importance of each scene is computed. According to our definition, the semantic importance indicates how well the scenes represent the complete sequence. One scene usually contains a few shots while some of these shots may be originated from the same camera. The shot segmentation is implemented according to the similarity of the two adjacent frames. In this work, the correlation coefficient of the luminance histograms is used to detect the shot boundaries, and the threshold is set as 0.9 for the cut change detection between the shots.

The semantic scene importance model is constructed based on the following considerations:

Audio genre and loudness: the audio classification is performed on the clips while the loudness is calculated for each frame. Three weights for the audio clips with different genres are set: 1 for speech and music, 0 for noise and silence, 0.5 for others. The average loudness \bar{L} over all the frames in a clip is computed first, and the product of the weight and \bar{L} is taken as the semantic audio importance of this clip. Finally, the average over all the clips in this scene is computed as the semantic audio importance (SAI) of this scene.

Scene representative index: The goal of video summarization is to find the most representative segments or frames. Hence, only unrelated shots should be included in the summary, whereas similar shots, which are usually shots with the same camera, should be excluded. A representative histogram of a shot is defined as the average of all frame histograms in this shot. All the representative histograms within one scene are compared using the correlation coefficients to classify shots which are close to each other into the same group of related shots and shots which differ from each other to different groups of unrelated shots. In this study, we use another threshold 0.8 to compare two representative histograms. Considering the importance for video summarization, we think that the shorter scene with more unrelated contents is more important than that longer scene with less different visual contents. Thus, the scene representative index (SRI) is defined as the ratio of the number of groups of related shots and the frame number in the scene.

Face & Text importance: For a general video, people are usually interested in face or text frames since they can help in understanding the semantics of the scene. We detect the dominant frontal face or text regions in this work, and the tiny face and text regions are not taken into account when computing the face & text importance. The size of a face or text region in a frame combined with the weight in Fig. 1 corresponding to its location is calculated as the importance index for this frame (FTF) [4]. Furthermore, the ratio of the detected face and text frames weighted by FTF compared to the scene length is computed as the face & text importance for this scene (FTS).

1/6	1/3	1/6
1/2	1	1/2
1/3	1/2	1/3

Fig.1 – Location weight for a frame image. The rectangle denotes a frame image, and numbers inside the rectangle are the weights for the corresponding location in the image.

Relationship with global emotion: usually, any movie can be classified into a special affective genre, e.g. comedy, thriller, tragedy, etc. The affective genre is closely connected to some audiovisual clues including pitch, loudness, luminance, and motion speed, called affective features. For example, luminance and pitch averages are usually high in comedy while low in tragedy [5][10]. Although it is a challenging issue to discriminate different emotions from audiovisual contents, we only try to use the distance between the affective features in one scene and the whole sequence to express the semantic relation of this scene to the overall sequence. If the emotion of one scene is close to the global emotion, we think this scene could represent the sequence in a summary. First, the pitch and loudness of every audio frame are computed, and the luminance and representative motion vector of every video frame are computed as well. Next, the averages of the affective features without the maximal and minimal 1% feature values are calculated and taken as the global affective features of the whole sequence. Furthermore, the averages of the affective features in each scene are defined as the affective scene features. Finally, the reciprocal of the Euclidean distance between the scene affective features and the global affective features is regarded as the relationship index (*RGE*) between the scene and the global emotion.

After deriving *SAI*, *SRI*, *FTS*, and *RGE*, these four values are normalized into the interval [0,1] over all the segmented scenes first. Based on the normalization of *SRI* and *FTS*, we define the semantic visual importance as following:

$$SVI = SRI + FTS \quad (2)$$

Then, adopting the ideas about the effect of audio and video contents on the joint influence in the audiovisual quality metric [11], we perform a fusion for semantic scene importance (*SSCI*) as following:

$$SSCI = SAI + SVI + RGE + SAI \cdot SVI \cdot RGE \quad (3)$$

C. Semantic shot importance

The construction of semantic shot importance model is similar but not identical to the scene importance model. Here the semantic audio importance (*SAI*) and face & text importance (*FTF*)

are also taken into account, while the processing unit is changed from a scene to a shot. Additionally, motion information (camera motion and local motion) and shot length are the other two factors which are considered.

Different types of camera motion are used to represent the objective reality for catching the attention of viewers. So, camera motion is a useful clue for semantic importance analysis, and it can be estimated according to the motion vectors in the background region. A shot usually contains a certain type of camera motion, such as panning, tilting, rolling, tracking, booming, dollying, zooming, and still. Each camera motion has its own expressive ability. For example, zooming/dollying is always employed to emphasize the details or an overview scene, and the emphasizing importance is proportional to the motion speed. On the other hand, panning or tilting usually reveals the surroundings horizontally or vertically while it cannot catch viewer attention in fast motion [12][13]. Although other camera motion types have also the respective expressive effect, it is difficult to map them to a single semantic purpose because of the scene complexity and director's subjective intention. In this work, only zooming/dollying and panning/tilting are taken into account. So, we make use of the following 4-parameter affine model to detect zooming/dollying, and panning/tilting, and the motion speed is approximated by the corresponding motion parameters [14].

$$\mathbf{MV}_{\text{camera}} = \begin{pmatrix} \text{zoom} & \text{rotate} \\ -\text{rotate} & \text{zoom} \end{pmatrix} \cdot \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} \text{pan} \\ \text{tilt} \end{pmatrix} \quad (4)$$

Although one shot usually contains one camera motion type, sometimes zooming/dollying and panning/tilting exist simultaneously in some special scenes. So, the camera motion importance *CMI* is defined as follows:

$$CMI = \begin{cases} 0, & \text{no zooming/dollying or panning/tilting} \\ \log(\max\{\text{zoom}, 1\}), & \text{only zooming/dollying} \\ \log(1 + \frac{1}{\max\{\text{pan}, \text{tilt}\}}), & \text{only panning/tilting} \\ \log(\max\{\frac{\text{zoom}}{\max\{\text{pan}, \text{tilt}\}}, 1\}), & \text{simultaneously} \end{cases} \quad (5)$$

Object motion, defined as the motion of the foreground object, is important information for video analysis, which can be obtained by the global motion compensation to motion vectors. To reduce computational complexity, we make use of motion vectors to represent the object motion. First, viewers generally pay more attention to fast motion, so the motion activity of each frame is defined as the average magnitude of all the motion vectors:

$$MA = \text{avg}(\|MV\|) \quad (6)$$

In addition, the temporal motion coherence is another useful clue for shot motion importance. The sensory data is processed in a continuous fashion, creating a continuous internal representation of the outside world [15]. Comparatively speaking, the continuous and homogeneous motion can attract more attention based on the observation on the videos with different motion types [13]. Thus, we define the entropy of motion vectors of the macro-blocks lying at the same location to express the temporal motion coherence of this location.

$$E_{TM} = -\sum_{n=1}^N h_k(n) \cdot \log(h_k(n)) \quad (7)$$

where h_k is the probability density function (PDF) of the phase histogram of motion vectors within the shot, and N is the number of the histogram bins. Then, the average of all the entropies $\text{avg}(E_{TM})$ is defined as the temporal motion coherence of this shot, and the importance index of the object motion is defined as follows:

$$OMI = \text{avg}(MA) / \text{avg}(E_{TM}) \quad (8)$$

where $\text{avg}(MA)$ is the average motion activity of this shot.

Additionally, as for the semantic shot importance in a scene, we think a longer shot is more important than a shorter one, because it can describe more story contents. So, the ratio of the shot length compared to the scene length is considered as another index for semantic shot importance, noted as SLR . Finally, similarly to the semantic scene importance model, all the semantic importance indices are normalized into the interval $[0,1]$ over all the shots in the scene, then the semantic visual importance (SVI) and the semantic shot importance ($SShI$) are defined as follows, respectively.

$$SVI = FTS + CMI + OMI + SLR \quad (9)$$

$$SShI = SAI + SVI + SAI \cdot SVI \quad (10)$$

D. Semantic frame importance

Because one isolated audio frame is semantically negligible for human perception, we only take visual information into account when computing semantic frame importance. The face & text importance (FTF) as well as the motion information are utilized. However, human faces are not present in all types of video content. If the video genre information is obtained from the Electronic Service Guide (ESG), the visual attention model proposed by Itti et al. [16] will

be employed to extract the attention region for those videos without human activity. The weighted size of face & text or attention regions is used as the region importance (RI), here the location weight matrix is identical to that in Fig. 1.

Furthermore, motion activity of the frame MA is still used whilst the spatial motion coherence is taken into account when constructing the semantic frame importance model. Similarly, we define the entropy of all the motion vectors in a frame to express the spatial motion coherence.

$$E_{SM} = -\sum_{n=1}^N h_k(n) \cdot \log(h_k(n)) \quad (11)$$

where h_k is the PDF of the phase histogram of motion vectors in the frame, and N is the number of the histogram bins. The semantic frame importance is defined as follows:

$$SFrI = RI + MA/E_{SM} \quad (12)$$

where RI and MA/E_{SM} for all frames are first normalized into the interval $[0,1]$.

III. VIDEO SUMMARIZATION BASED ON SEMANTIC IMPORTANCE

As its name implies, a video summary is much shorter than the original video, and users may have a preferred summary duration. On the other hand, we think a scene is more important than a shot, which in turn is more important than a frame, because one single frame cannot express enough information for video understanding. We thus propose the following summarization approach (in this study, the video summary is a collection of a set of static key frames, and we assume that the number of the key frames is predefined to L):

Step 1: because a video summary is much shorter than the original video, we assume to drop half of all the scenes with the lower semantic importance from the summary when extracting the key frames.

Step 2: for each scene within the remaining scenes, compute the proportion ($p1$) of its semantic importance compared to the sum of the importance of all the scene candidates. Then, the number of key frames of this scene is set as $ScL = [p1 \cdot L]$, where $[\cdot]$ denotes the rounding operation.

Step 3: for a set of shots containing similar visual contents within each selected scene, only the shot having the highest $SShI$ (in this case, the highest SAI usually corresponds to the highest $SShI$) is selected for the extraction of key frames. Therefore, we can get a few representative shots with different visual contents. For every selected shot, compute the proportion ($p2$) of its semantic importance compared to the sum of the importance indices of all

the selected shots. Then, the number of key frames of this shot is set to $SpL = \lceil p2 \cdot ScL \rceil$.

Step 4: within each shot, the key frames are selected in terms of the order of the semantic frame importance, till the number of key frames of this shot achieves SpL .

IV. EXPERIMENTS AND EVALUATION

To evaluate the proposed semantic audiovisual analysis approach for video summarization, four typical video genres were tested: news, sports, nature documentary, and comedy. The duration of each video was 3 minutes. Fig. 2 gives the results of the semantic importance analysis of the comedy sequence for scenes, shots, and frames, as well as the extracted key frames in two adjacent shots. The horizontal axes in (a), (b), (c), and (d) denote the frame index while the scales are different to each other. The horizontal axes denote the frame index of the whole sequence in (a) and (b), while it is the frame index in the scene in (c), as well as the frame index in the shots in (d).

In addition, the key frames were extracted from the original sequences according to the approaches in Section 3 with the predefined number of key frames equal to 5% of the frames of the original video. To evaluate the expressive ability of the summarization results, 10 untrained volunteers were asked to watch the original videos firstly, and then the key frames. The key frames were shown as an image set, lasting as long as the volunteers thought they had understood the meaning of these images. After each video, a questionnaire about the expressive ability of the summarization was completed by each volunteer based on his/her own subjective judgment. The expressive ability is divided into 5 levels from 1 to 5 corresponding to the worst to the best expression, and then the average of 10 volunteers' assessment is computed and listed in Table 1. As a comparison, we

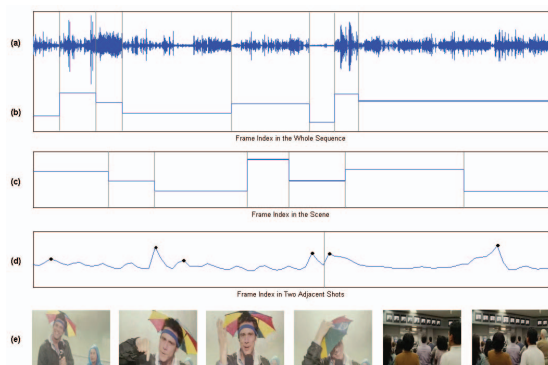


Fig. 2 – Analysis results for the comedy sequence. (a) audio samples; (b) semantic scene importance (gray vertical lines denote the scene boundaries); (c) semantic shot importance (gray vertical lines denote the shot boundaries); (d) semantic frame importance (gray vertical line denotes the shot boundary, and the dots denote the positions of the extracted key frames); (e) extracted key frames in two adjacent shots.

also give the experimental results using the algorithms in [3] and [4], where the results of news, sports, and nature documentary sequences were obtained in our earlier experiments in [4], while the results of the comedy sequences were obtained from the current experiment.

Table 1. Evaluation of video summarization

Video genre	news	sports	natural document.	comedy
Proposed	3.5	3.7	3.3	3.5
algorithm in [3]	3.0	2.3	2.5	2.1
algorithm in [4]	3.4	3.1	3.0	3.2

Based on the experimental results, the proposed semantic analysis approach is effective for extracting the static key frames. In addition, according to the semantic importance analysis for scenes and shots, it is easy to extend the static key frames to the dynamic summarization which can be composed of the key audiovisual clips with high semantic importance.

V. CONCLUSION

An effective semantic audiovisual analysis approach for video summarization was presented in this study. The sequence is first segmented into scenes based on selected audio characteristics, and then scenes are divided into shot groups with unrelated visual contents based on luminance histograms. The semantic importance is analyzed and calculated hierarchically from each scene, shot, and video frame. Selected audio and video features are chosen for the calculation of the semantic importance. The approach aims at extracting the most representative scenes, shots, and frames whilst the similar key frames are eliminated from the video summary. The subjective evaluation results demonstrated the promising performance of the proposed semantic analysis approach for video summarization.

REFERENCES

- [1]. M. M. Yeung, and B-L Yeo, "Video Content Characterization and Compaction for Digital Library Application," in Proc. SPIE'97, Storage and Retrieval of Image and Video Database V, San Jose CA, Feb. 1997.
- [2]. H. Sundaram, and S-F Chang, "Computable Scenes and Structure in Films," *IEEE Trans. Multimedia*, vol. 4, no. 4, pp. 482-491, Dec. 2002.
- [3]. Y-F Ma, X-S Hua, L. Lu, and H-J Zhang, "A Generic Framework of User Attention Model and Its Application in Video Summarization," *IEEE Trans. Multimedia*, vol. 7, no. 5, pp. 907-919, Oct. 2005.
- [4]. J. You, G. Liu, L. Sun, and H. Li, "A Multiple Visual

Models Based Perceptive Analysis Framework for Multilevel Video Summarization,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 3, pp. 273-285, Mar. 2007.

- [5]. A. Hanjalic, and L-Q Xu, “Affective Video Content Representation and Modeling,” *IEEE Trans. Multimedia*, vol. 7, no. 1, pp. 143-154, Feb. 2005.
- [6]. D. Tjondronegoro, Y-P P. Chen, and B. Pham, “Highlights for more complete sports video summarization,” *IEEE Multimedia*, vol. 11, no. 4, pp. 22-37, Oct.-Dec. 2004.
- [7]. Y. Wang, Z. Liu, and J-C Huang, “Multimedia Content Analysis using Both Audio and Visual Clues,” *IEEE Signal Processing Mag.*, vol. 17, no. 6, pp. 12-36, Nov. 2000.
- [8]. D. Li, I. K. Sethi, N. Dimitrova, and T. McGee, “Classification of general audio data for content-based retrieval,” *Pattern Recognition Letter*, vol. 22, no. 5, pp. 533-544, Apr. 2001.
- [9]. Z. Liu, Y. Wang, and T. Chen, “Audio feature extraction and analysis for scene segmentation and classification,” *Journal VLSI Signal Processing*, vol. 20, pp. 61-79, 1998.
- [10]. R. Picard, *Affective Computing*, Cambridge, MA: MIT Press, 1997.
- [11]. J. G. Beerends, and F. E. D. Caluwe, “The influence of video quality on perceived audio quality and vice versa,” *Journal Audio Eng. Soc.*, vol. 47, no. 5, pp. 355-362, May 1999.
- [12]. Y-F Ma, L. Lu, H-J Zhang, and M. Li, “A user attention model for video summarization,” in *Proc. 10th ACM Int. Conf. Multimedia*, Juan-les-Pins, France, Dec. 1-6, 2002, pp. 533-542.
- [13]. Z-C Zhao, A-N Cai, “Extraction of semantic keyframes based on visual attention and affective models,” *IEEE Int. Conf. Computational Intelligence, Security*, Harbin, China, Dec. 15-19, 2007, pp. 371-375.
- [14]. R. Wang, and T. Huang, “Fast Camera Motion Analysis in MPEG domain,” in *Proc. IEEE Int. Conf. Image Processing*, Kobe, Japan, Oct. 24-28, 1999, pp.691-694.
- [15]. T. J. Smith, “An attentional theory of continuity editing,” Ph. D thesis, University of Edinburgh, 2005.
- [16]. L. Itti, C. Koch, and E. Niebur, “A Model of Saliency-Based Visual Attention for Rapid Scene Analysis,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254-1259, Nov. 1998.

Junyong You (M’08) received his PhD degree in Electronics and Information Engineering from Xi’an Jiaotong University, Xi’an, China, in 2007. Dr. You is currently a senior researcher with Tampere University of Technology, Tampere, Finland. His research interests include objective audiovisual quality metrics, audiovisual content analysis, video coding, wavelet analysis and application, image processing, and non-stationary signal analysis.

Miska M. Hannuksela (M’02) received the MS degree in engineering from Tampere University of Technology, Tampere, Finland, in 1997. He is currently a Research Leader in Nokia Research Center, Tampere, Finland. From 1996 to 1999, he was a Research Engineer with Nokia Research Center in the area of mobile video communications. From 2000 to 2003, he was a Project Team Leader and Specialist in various mobile multimedia research

and product projects in Nokia Mobile Phones. Since 2003, he has been a Research Manager, Senior Research Manager, and Research Leader heading teams in the area of visual technologies and real-time multimedia communications in Nokia Research Center. He has been an active participant in the ITU-T Video Coding Experts Group since 1999 and in the Joint Video Team of ITU-T and ISO/IEC since its foundation in 2001. He has also contributed to several other multimedia standards, such as IP data casting over DVB-H and 3GPP multimedia services. His research interests include video error resilience, scalable video coding, and video communication systems. He has co-authored several tens of papers in these fields.

Moncef Gabbouj (SM’95) received his BS degree in electrical engineering in 1985 from Oklahoma State University, Stillwater, and his M.S. and PhD degrees in electrical engineering from Purdue University, West Lafayette, Indiana, in 1986 and 1989, respectively. Dr. Gabbouj is currently a Professor at the Department of Signal Processing at Tampere University of Technology, Tampere, Finland. He was Head of the Department during 2002-2007. His research interests include multimedia content-based analysis, indexing and retrieval; nonlinear signal and image processing and analysis; and video processing and coding. Dr. Gabbouj served as Distinguished Lecturer for the IEEE Circuits and Systems Society in 2004-2005. He served as associate editor of the IEEE Transactions on Image Processing, and was guest editor of Multimedia Tools and Applications, the European journal Applied Signal Processing. He is the past chairman of the IEEE Finland Section, the IEEE CAS Society, Technical Committee on DSP, and the IEEE SP/CAS Finland Chapter. Dr. Gabbouj was the recipient of the 2005 Nokia Foundation Recognition Award and co-recipient of the Myril B. Reed Best Paper Award from the 32nd Midwest Symposium on Circuits and Systems and co-recipient of the NORSIG 94 Best Paper Award from the 1994 Nordic Signal Processing Symposium. He is a member of IEEE SP and CAS societies.