# Strategies for Development and Evaluation of Video Summarization Algorithms

Marcos Vinicius Mussel Cirne[1], Helio Pedrini
Institute of Computing, University of Campinas
Campinas-SP, Brazil, 13083-852

***Abstract*—Video summarization is a challenging field of research which consists of generating a synopsis of a given video containing the most important events. It is also suitable for analyzing large amounts of digital videos, being helpful on speeding up the tasks of indexing, browsing and content-based retrieval. This work presents three different approaches that deal with videos of any genre, along with their respective evaluation metrics and evolutions between these approaches. Results show that it was possible to develop a method that not only has a superior quality than the state-of-art but also is very fast and efficient, being applicable in video management environments.**

***Keywords*-video summarization; content analysis; frame dissimilarity.**

## I. Introduction

With the advent of the newest technologies, it has become much easier and more accessible for people to record high quality videos with their digital cameras, smartphones or tablets. Aside from that, the growth of video hosting websites (including cloud platforms), social networks and video streaming services compels their respective users to upload and share a huge number of videos. This demands a great effort for these systems to store them in such a way the tasks of video indexing and retrieval become efficient enough to provide an adequate service for the users.

Much research has been done in order to develop techniques that are capable of manipulating these data in an automatic, efficient and accurate way, concerning the issues of searching, browsing, retrieval and content analysis. Among these techniques, there is video summarization [1], [2], [3], [4], [5], which analyzes the content of a given video and creates a snippet that preserves the most important information of this video. However, this process must be conducted in such a way that, by watching the summary, the users must be able to understand at least the most part of the original content without needing to turn to the original video. Moreover, defining what is relevant or not in video summarization is an open problem, since a content that is important to ones may not be to others, besides the fact that there are several categories of videos, such as sports, documentaries, talk shows, home videos, among others.

This work investigates the principles of video summarization and the most common strategies used in the stages of the summarization process, as well as the evaluation metrics and databases used to validate a method. Results of each method are then compared against several video summarization methods both in terms of quality of produced summaries and similarity to human-produced summaries. The main contribution of this work is the development of a video summarization method that works with any video genre, also having a superior quality in relation to similar methods and being applicable to frameworks that deal with video indexing, retrieval and browsing.

This paper is organized as follows: Section II reviews some video summarization methods of the literature; Section III describes each of the methods proposed for video summarization; Section IV presents results obtained by each proposed method, also comparing to other related methods; Section V makes some conclusions about the work.

## II. Literature Review

Several video summarization techniques have been developed with the goal of generating summaries that reflect a "common sense", i.e., which encompass the contents that most of humans would select for a summary. Some of them focus on a specific genre, which makes it easier to define the criteria for choosing the most important events of all videos from that genre. Usually, this type of approach produces very accurate results, but in expense of the genre constraint. On the other hand, there are also approaches that work with any kind of content, but the results are less accurate and demand more generic ways to describe the features that will be used for the summarization process.

To generate a summary, one may search for specific events that frequently occur on videos of the same content, such as goals or attack situations in soccer games, climax scenes in movies or even contents that appear between parts of a TV show, such as advertisings. These elements are known as high-level features, which are a semantic representation of a content that happen on a given instant of the video and refers to subjective aspects. Other elements that fit in this category are: time, space, objects and human actions. Another possibility is to observe details of the video frames which do not represent directly a semantic element, but are useful for analyzing repeatable patterns or how an object moves in a sequence of frames. Those are called low-level features, and they include: color histograms, texture, motion, audio, subtitles, etc.

Concerning the existing video summarization methods, Mundur et al. [6], developed an approach which uses the

[1] Ph.D. Thesis.

Delaunay Triangulation (DT) algorithm to cluster video frames and picks the centroids of each cluster to compose the summaries; Furini et al. [7] proposed STIMO (STIll and MOving Video Storyboards), which generates both static and dynamic summaries using HSV color histograms, clustering frames through a variation of the Farthest Point-First (FPF) algorithm [8]; Avila et al. [9] proposed a very simple approach for video summarization named VSUMM, that uses $K$-means for clustering and HSV histograms for color features; Almeida et al. [10] presented an approach named VISON (VIdeo Summarization for ONline applications), which summarizes videos by working directly on the compressed domain and allows user interaction to control the quality of the summaries; Mahmoud et al. [11] proposed VSCAN, a summarization method that uses the DBSCAN clustering algorithm and describes the frames by combining color and text features.

## III. Methodology

This section describes three different approaches to video summarization. The first one [2] is based on spectral clustering algorithms, defined as a method that clusters data points using eigenvectors of matrices calculated from a given dataset and usually outperforms traditional clustering algorithms, such as $K$-means. Concerning the video summarization context, it can be used in several tasks, including keyframe extraction, shot boundary detection and important events detection.

Initially, given a video, the method samples it in a smaller number of frames (5 frames per second) in order to increase performance. Then, it passes by a feature extraction stage, where feature vectors are extracted from keypoints detected on each sampled frame by means of specific algorithms such as SIFT (Scale-Invariant Feature Transform) [12], SURF (Speeded-Up Robust Features) [13] and ORB (Oriented FAST and Rotated BRIEF) [14]. Next, the shot boundaries are detected by analyzing a visual rhythm by histogram image computed from all the sampled frames. This step gives an estimation of the number $k$ of shots, which will be used as a reference to extract the frames that will make part of the final summary (one frame per shot). The spectral clustering algorithm is started by constructing an affinity matrix of size $n \times n$ (where $n$ is the number of sampled frames) such that each position $A(i, j)$ stores the distance of frame $i$ to frame $j$. Then, from a normalized matrix of size $n \times k$ calculated from the $k$ largest eigenvectors of $A$, its rows are then clustered in $k$ groups. Finally, if a row $i$ of that matrix was assigned to cluster $j$, the associated frame $i$ will also be assigned to that cluster. From the computed clusters, the key frames are extracted by taking the frames that are closest to each respective centroid. At the last step, key frames that have similar contents to at least one other are discarded. The remainder frames will then comprise the summary.

The advantage of this method is that every stage is executed in an unsupervised fashion, such that the number of shots does not need to be known *a priori*. However, the whole summarization process is still expensive, because of the spectral clustering, even though it leads to more accurate results than standard clustering approaches. In addition, the distance metric used to find similarities in frames is slow, since it compares the matches between all pairs of keypoints between two given frames.

To overcome those issues, the second method, named VSQUAL [3], adopted an approach based on objective Image Quality Assessment (IQA) metrics [15], [16]. Their main characteristic is the exploitation of physiological and psychophysical characteristics of the human visual system (HVS), at the same time they take into account the structural information of images. On the other hand, they have high sensitivity to geometric changes, such as translation, scaling, rotation, and so on. For this work, the FSIM metric (Feature Similarity Index) [16] was used as the similarity measure, which was designed from the principle that the HVS interprets a scene by analyzing the information contained in salient low-level features, such as edges and zero-crossings. Unlike the majority of IQA metrics, FSIM can be easily extended to work with color images, leading to a new measure called $\text{FSIM}_C$. Since color information is a fundamental part of image and scene understanding, it is expected that FSIM performs better than other approaches.

With respect to the methodology pipeline, the frame sampling step is the same as done in the previous method. After that, a similarity matrix $A$ is constructed such that $A(i, j) = \text{FSIM}_C(i, j)$, where each value ranges from 0 (no similarity) to 1 (full similarity). At the shot boundary detection step, The $\text{FSIM}_C$ values between consecutive frames are analyzed. A transition is detected by analyzing subsets of size $m = 9$ of $\text{FSIM}_C$ values between $m$ pairs of consecutive frames. If the middle value of a subset is the minimum and the difference between the maximum and the minimum value is above a threshold (empirically defined at 0.1), a transition is detected. One key frame is extracted from each shot, based on submatrices of $A$ related to the frames contained in the respective shots. A redundancy elimination process is run by comparing the $\text{FSIM}_C$ values between all pairs of key frames, and whenever a value exceeds a similarity threshold (defined at 0.75), a key frame is discarded for the final summary.

The advantage of VSQUAL resides on the fact that the video frames are represented by a more objective measure, which reflects the way that humans perceive images. Moreover, comparing to the spectral clustering method, VSQUAL does not employ clustering algorithms, extracting keyframes in a simpler way, and it can be easily adapted to any other image quality metric. However, this approach still has some problems in dealing with videos that have considerable movement, which causes oscillations at the $\text{FSIM}_C$'s between consecutive frames.

To solve the aforementioned problems, as well as to improve the general performance of the whole summarization process, a third method, called VISCOM (VIdeo Summarization by Color Co-Occurrence Matrices), was developed [17]. This method extends the previous one by using color co-occurrence matrices (CCM) [18], [19], [20] as the image descriptor for video frames. These matrices are commonly used to represent the distribution of color features between pairs of pixels in an image, considering the correlations between the color bands as well. Figure 1 shows an example for the RGB color space.

The construction of the CCM's from a multispectral image $I$ goes as follows: let $C_1, C_2, ..., C_n$ be the $n$ channels of $I$, where each one is coded on $L$ levels, and $L$ the number of rows and columns of the CCM's. Also, let $C_u$ and $C_v$ be a pair of channels (with $1 \leq u, v \leq n$). Finally, let $p = (x, y)$, where $0 \leq x \leq H-1$ and $0 \leq y \leq W-1$, be a pixel in $I$ and $q = (\Delta x, \Delta y)$, with $\Delta x = x + d \ \cos\theta$ and $\Delta y = y + d \ \sin\theta$, a translation of $p$, such that $q$ remains in the spatial domain of $I$. The computation of each position $(i, j)$ of the CCM of size $L \times L$ and a translation vector $t$, for a pair of channels $C_u$ and $C_v$, is done according to Equation 1:

$$\text{CCM}_{(C_u,C_v)}(i, j \mid t) = \text{card}\{\{p, q\} \in I \quad \text{such that} \\ C_u(p) = i, C_v(q) = j\} \tag{1}$$

where $i$ and $j$ range from 1 to $l$.

For VISCOM, the RGB color space was used to represent the video frames, with $l = 8$ and $t = (1, 0)$ (one pixel to the right). Since $\text{CCM}_{t,(C_u,C_v)}$ and $\text{CCM}_{t,(C_v,C_u)}$ store the same information, there are only 6 possible pairs of channels $(C_u, C_v)$, thus leading to 6 different CCM's: (R,R), (R,G), (R,B), (G,G), (G,B) and (B,B).

To measure the distance between pairs of frames, the Normalized Sum of Square Differences (NSSD) [21] is used, which has been proved to be very robust and widely used in tasks that deal with digital image correlation [22]. This function is defined according to Equation 2:

$$\text{NSSD}(I_c, J_c) \;=\; \frac{\displaystyle\sum_{i=1}^{l}\sum_{j=1}^{l}(I_c(i,j) - J_c(i,j))^2}{\sqrt{\displaystyle\sum_{i=1}^{l}\sum_{j=1}^{l}(I_c(i,j))^2 \times \sum_{i=1}^{l}\sum_{j=1}^{l}(J_c(i,j))^2}} \tag{2}$$

where $l$ is the matrix size, $c$ is one of the six co-occurrence matrices described earlier, whereas $I_c$ and $J_c$ are two co-occurrence matrices that represent different images. NSSD ranges from 0 to 1, where the closer to zero, the more similar are the images.

Figure 2 shows the methodology pipeline for VISCOM. At the first stage, the frames are sampled in a smaller amount in order to save some computational time for the whole summarization process, at the same time it does not discard any piece of meaningful information. In this work, a sampling of 15 frames (defined by means of empirical tests) was used, i.e., it takes the first frame of the original video, along with the 16th, the 31st, and so on. In other words, for a video with a frame rate of 30 frames per second, two frames are extracted per second of video.

Once the frames are sampled, the CCM's are computed for every sampled frame. Then, the NSSD's between consecutive frames are computed, rather than computing for every possible pair. These computed values are used for the shot boundary detection stage in a similar way to VSQUAL. After the shots are detected, the middle core frame of each shot is regarded as the representative frame, resulting in a set of key frames. At the last step, a redundancy elimination algorithm is executed by analyzing the NSSD's between all pairs of keyframes. If a NSSD is below a distance threshold (empirically defined as 0.2), one frame is discarded for the final summary.

The advantage of VISCOM lies on the robustness and effectiveness of both the image descriptor and the distance function, aiding the whole method in achieving a reasonable performance. On the other hand, regarding the identification of similarities in images, it can still find some false positives and negatives, in a sense that pairs of images that have different contents but similar color distributions may lead to low values for the distance function or vice-versa.

## IV. Results

The tests were conducted on an AMD FX-6300 3.5 GHz processor and 4 GB of memory. All methods described in Section III were implemented with the OpenCV platform[1]. A collection of 50 videos of several genres from Open Video Project (OVP)[2] was used in the experiments. Together, all of the video sequences have a total duration of approximately 75 minutes (with each video lasting between 1 and 4 minutes) and 150,000 frames, whose original dimensions are 352 × 240 pixels. Due to space limitations, only the results for VISCOM will be shown, since it produced summaries with more quality than the previous two methods. In order to measure the performance of VISCOM, the implementation was executed 10 times with all videos from the database. The average execution time was 606.7 ± 2.8 seconds (about 12 seconds per video). These times are very satisfactory, once each summary is generated in a small percentage of the total time of each respective video (usually between 10% and 20%). Such performance is affected not only by the video frame count, but also by the number of extracted keyframes, because the higher this number is, more comparisons are made during the redundancy elimination stage.

To evaluate the summaries, a modified version of the CUS metric [9] was used. In this metric, for each video, the automatic summaries of each method are compared to manual summaries produced by 5 different users (ground-truth). If a pair of frames (one from an automatic summary and other from an user summary) is considered similar, these frames are removed from the next iteration of CUS. The concept of frame similarity is the same used in the redundancy elimination step described in Section III, where two frames are considered similar if the NSSD between them is equal to or greater than $T_S = 0.2$.

The score computation is based on three different values: number of similar frames $SF_i$ (which corresponds to frames from automatic summary that match frames from user summaries), number of frames in the automatic summaries $AS_i$

[1] http://www.opencv.org

[2] http://www.open-video.org

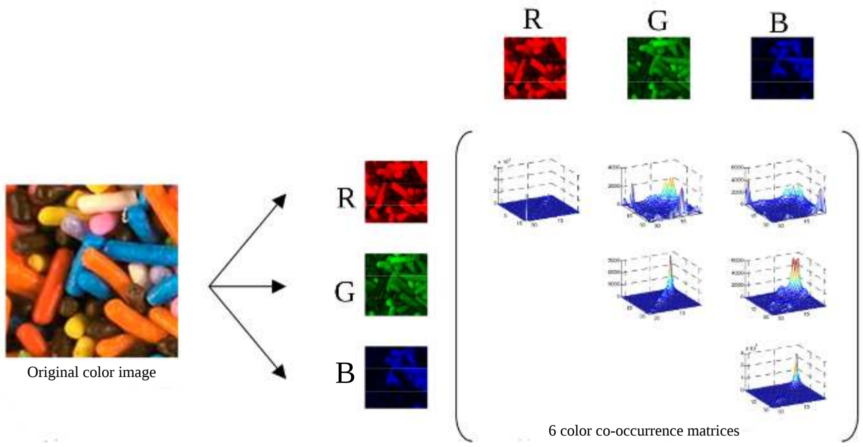

Fig. 1. Color co-occurrence matrices extracted from a image in the RGB color space. Image extracted from [18].

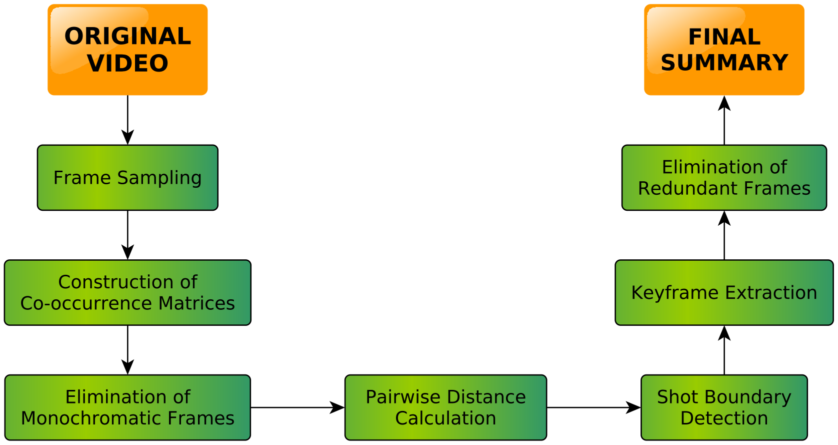
ORIGINAL VIDEO
Frame Sampling
Construction of Co-occurrence Matrices
Elimination of Monochromatic Frames
Pairwise Distance Calculation
Shot Boundary Detection
Keyframe Extraction
Elimination of Redundant Frames
FINAL SUMMARY

Fig. 2. Overview of the stages of VISCOM method.

and number of frames in the user summaries $US_i$, where $i \in \{1, 2, 3, 4, 5\}$ represents a specific user. From these values, both precision $P_i$ and recall $R_i$ values can be obtained, with $P_i = SF_i \ / \ AS_i$ and $R_i = SF_i \ / \ US_i$. Since there is a trade-off between precision and recall [10], the F-measure is used as the quality assessment metric for the automatic summaries. For each video, the F-measure is obtained by the average of the harmonic means of $P_i$ and $R_i$. Table I shows the average precisions, recalls and F-measures for the summaries of all videos from the database for each method. It can be seen, from the table, that VISCOM overcomes the evaluated state-of-the-art approaches, producing competitive results while maintaining a good trade-off between speed and quality.

TABLE I
AVERAGE PRECISION, RECALL AND F-MEASURES OF THE SUMMARIES PRODUCED BY EACH METHOD FOR THE ENTIRE DATABASE.

| Method | Precision (%) | Recall (%) | F-Measure |
|---|---|---|---|
| DT [6] | 54.7 | 43.3 | 0.469 |
| STIMO [7] | 51.9 | 62.1 | 0.552 |
| OVP | 58.4 | 65.7 | 0.589 |
| VSQUAL [3] | 55.7 | 74.3 | 0.608 |
| VISON [10] | 59.5 | 67.5 | 0.619 |
| VSUMM [9] | 72.1 | 64.1 | 0.666 |
| VSCAN [11] | 62.5 | 83.1 | 0.702 |
| **VISCOM** [17] | **64.9** | **81.1** | **0.706** |

Figure 3 shows the results of VISCOM for a specific video, along with the results generated from other summarization methods and their respective F-measures, as well as the summaries that were manually made by 5 different users. In this example, the automatic summaries covered all the content selected by the users. However, the key difference between the summaries lies on the size of the produced summaries related to the average size of the user summaries. Both VISCOM and VISON performed very well in this task, but the former was a slightly better in terms of F-measure. The other approaches either managed to cover the whole content, but in expense of a higher size of the final summary (STIMO, OVP and VSCAN), or produced shorter summaries with a less satisfactory content (DT, VSQUAL and VSUMM). The similar happens in the example shown in Figure 4, where VISCOM, VSUMM and OVP achieved the best results, but the other approaches obtained worse results because of the size of their respective summaries.

It is worth mentioning that, when comparing a frame from an automatic summary and one from a user summary, even though these frames belong to a same shot in the original video and have similar contents, false negatives can still be obtained due to large offsets between the frames, which lowers the F-measure values. VISCOM chooses the middle core frame of the detected shots, whereas VSUMM and VSCAN, for example, use clustering algorithms to group the frames in shots (clusters), selecting the ones that are closest to the centroids of each cluster as keyframes. However, any change in the keyframe selection strategy might cause a significant increase in the computational time, at the same time it is not worth the eventual gain on the average F-measure, since the visual changes in the summaries are little.

Despite the aforementioned issues, the NSSD function is very helpful in the task of identifying image similarities for the absolute majority of the cases. In addition, the keyframe selection strategy used in VISCOM is very suitable for generating satisfactory summaries that reflect the humans' concept of importance. Any change in this strategy might cause a significant increase in the computational time, at the same time it is not worth the eventual gain on the average F-measure, since the visual changes in the summaries are little.

## V. Conclusions

This thesis investigated the field of video summarization, providing a general contextualization, as well as an analysis of different methods of the literature in terms of knowledge domain, strategies for each stage of the summary creation process and metrics that assess the quality of the summaries. Three different approaches were proposed, differing in several aspects, such as how the video frames were described, how the keyframes were selected to compose the final summary and the metrics used to identify image similarities when comparing to a ground-truth. The summaries generated by each method covered most aspects of the contents of each video used in the tests, producing satisfying results.

Additionally to the papers [2], [3], [17] related to the three developed approaches, an image clustering method based on Partial Least Squares [23] was also proposed and applied to the video summarization problem. This work, presented in SIBGRAPI'2015, received an invitation to be extended for a special issue at Pattern Recognition journal, which is currently under review [24].

Some directions of future work include: analysis of some variations for the construction of color co-occurrence matrices, tuning of the parameters used in some stages of VISCOM's pipeline and combination of other features to the image descriptor, such as motion and spatio-temporal features. Furthermore, a deeper analysis of the distance function for detecting similarities in pairs of images can also be considered, with the objective of reducing the number of both false positives and negatives and, therefore, leading to more accurate evaluations of summaries.

## References

[1] M. Ajmal, M. H. Ashraf, M. Shakir, Y. Abbas, and F. A. Shah, "Video Summarization: Techniques and Classification," in *International Conference on Computer Vision and Graphics*. Warsaw, Poland: Springer-Verlag, 2012, pp. 1–13.

[2] M. V. M. Cirne and H. Pedrini, "Video Summarization Method Based on Spectral Clustering," in *18th Iberoamerican Congress on Pattern Recognition*, vol. 8259, Havana, Cuba, 2013, pp. 479–486.

[3] ——, "Summarization of Videos by Image Quality Assessment," in *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, ser. Lecture Notes in Computer Science, 2014, vol. 8827, pp. 901–908.

[4] Q. Luan, M. Song, C. Y. Liau, J. Bu, Z. Liu, and M.-T. Sun, "Video Summarization based on Nonnegative Linear Reconstruction," in *IEEE International Conference on Multimedia and Expo*, Chengdu, China, Jul. 2014, pp. 1–6.

[5] S. Mei, G. Guan, Z. Wang, S. Wan, M. He, and D. D. Feng, "Video Summarization via Minimum Sparse Reconstruction," *Pattern Recognition*, vol. 48, no. 2, pp. 522–533, 2015.

[6] P. Mundur, Y. Rao, and Y. Yesha, "Keyframe-Based Video Summarization Using Delaunay Clustering," *International Journal on Digital Libraries*, vol. 6, pp. 219–232, Apr. 2006.

[7] M. Furini, F. Geraci, M. Montangero, and M. Pellegrini, "STIMO: STIll and MOving Video Storyboard For The Web Scenario," in *Multimedia Tools and Applications*, vol. 46. Hingham, MA, USA: Kluwer Academic Publishers, 2010, pp. 47–69.

[8] T. F. Gonzalez, "Clustering to Minimize the Maximum Intercluster Distance," *Theoretical Computer Science*, vol. 38, pp. 293–306, 1985.

[9] S. E. F. de Avila, A. P. B. Lopes, A. da Luz Jr., and A. de Albuquerque Araújo, "VSUMM: A Mechanism Designed to Produce Static Video Summaries and a Novel Evaluation Method," *Pattern Recognition Letters*, vol. 32, no. 1, pp. 56–68, 2011.

[10] J. Almeida, N. J. Leite, and R. S. Torres, "VISON: VIdeo Summarization for ONline Applications," *Pattern Recognition Letters*, vol. 33, no. 4, pp. 397–409, Mar. 2012.

[11] K. M. Mahmoud, M. A. Ismail, and N. M. Ghanem, "VSCAN: An Enhanced Video Summarization Using Density-Based Spatial Clustering," in *Lecture Notes in Computer Science*, vol. 8156. Springer, 2013, pp. 733–742.

[12] D. Lowe, "Object Recognition from Local Scale-Invariant Features," in *Seventh IEEE International Conference on Computer Vision*, vol. 2, 1999, pp. 1150–1157.

[13] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded Up Robust Features," in *9th European Conference on Computer Vision*, May 2006, pp. 404–417.

[14] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An Efficient Alternative to SIFT or SURF," in *IEEE International Conference on Computer Vision*, Barcelona, Spain, 2011.

[15] S. Chikkerur, V. Sundaram, M. Reisslein, and L. J. Karam, "Objective Video Quality Assessment Methods: A Classification, Review, and Performance Comparison," *IEEE Transactions on Broadcasting*, vol. 57, no. 2, pp. 165–182, Jun. 2011.

[16] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: A Feature Similarity Index for Image Quality Assessment," *IEEE Transactions on Image Processing*, vol. 20, no. 8, pp. 2378–2386, Aug. 2011.

[17] M. V. M. Cirne and H. Pedrini, "VISCOM: A Fast and Robust Video Summarization Approach Using Color Co-occurrence Matrices," *Multimedia Tools and Applications (under review)*, 2015.

[18] V. Arvis, C. Debain, M. Berducat, and A. Benassi, "Generalization of the Coocurrence Matrix for Colour Images: Application to Colour Texture Classification," *Image Analysis & Stereology*, vol. 23, no. 1, pp. 63–72, 2011.

[19] M. B. Islam, K. Kundu, and A. Ahmed, "Texture Feature Based Image Retrieval Algorithms," *International Journal of Engineering and Technical Research*, vol. 2, pp. 170–173, Apr. 2014.

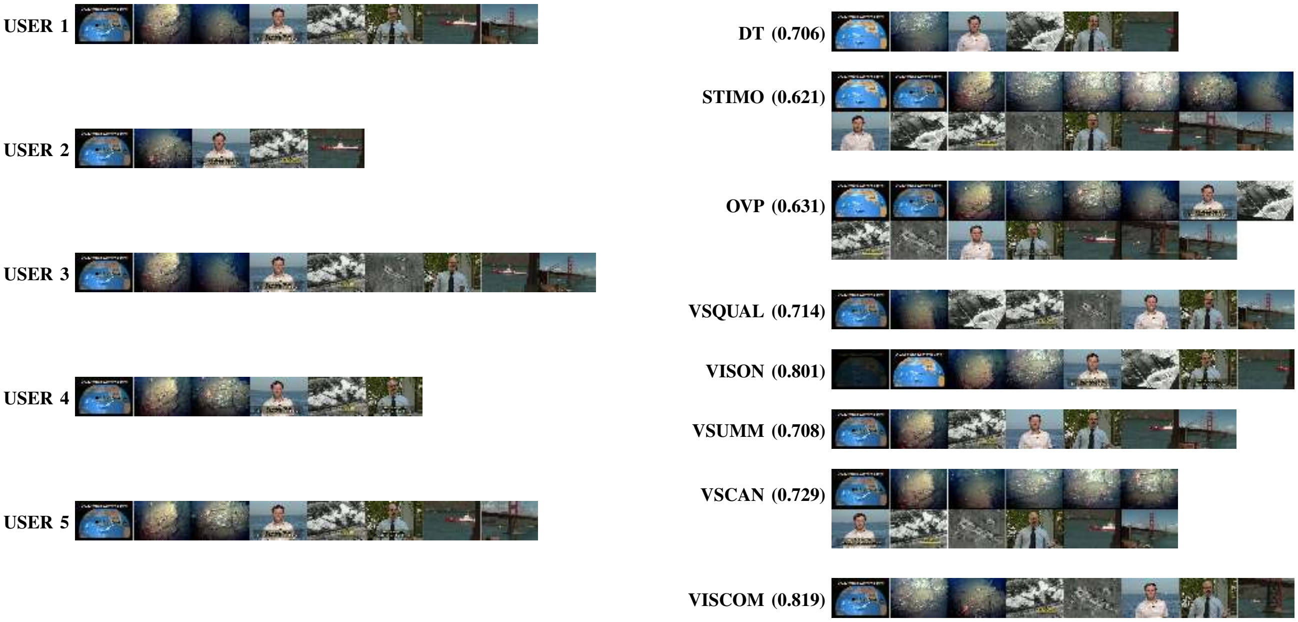

Fig. 3. User summaries and automatic summaries of each method from the video *America's New Frontier, Segment 10*, along with the respective F-measures.

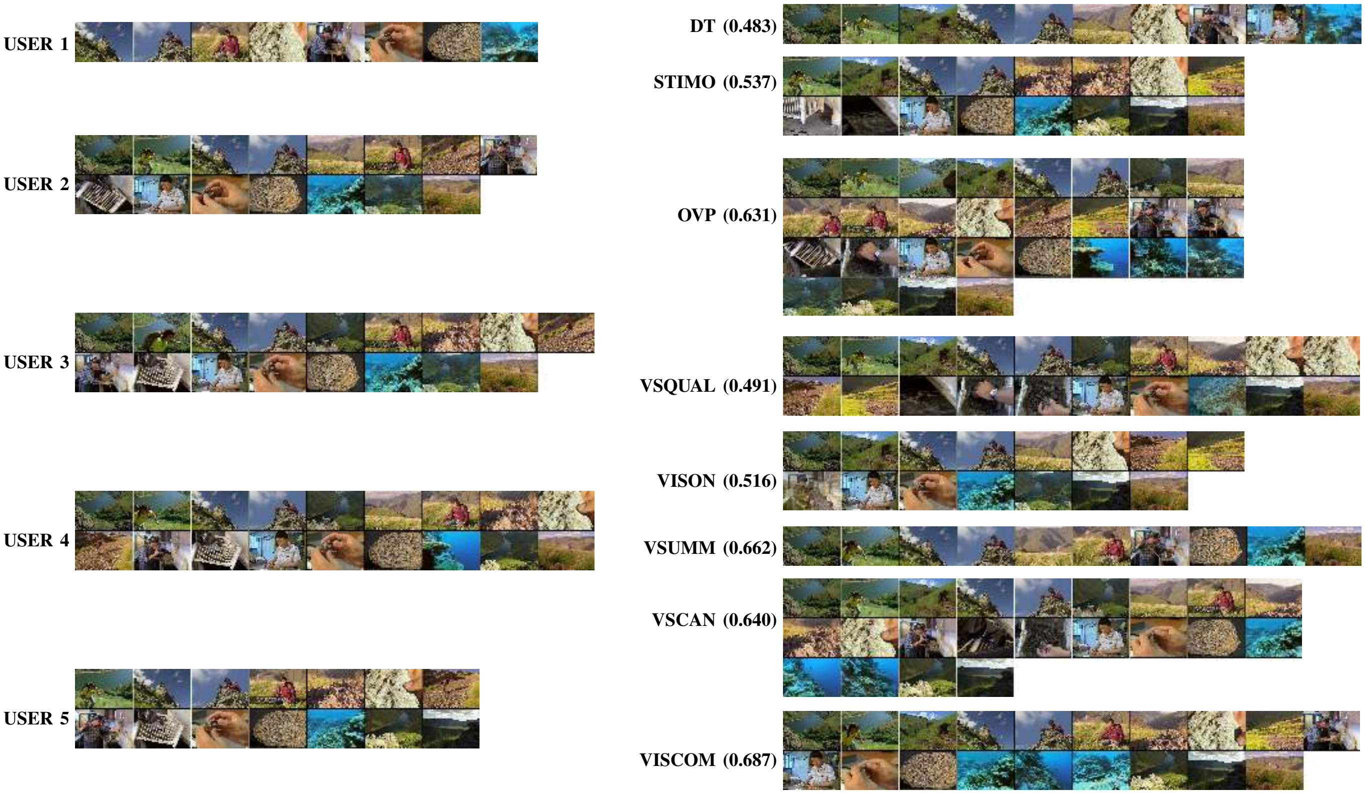

Fig. 4. User summaries and automatic summaries of each method from the video *Exotic Terrane, Segment 04*, along with the respective F-measures.

[20] A. L. Lavanya and R. Sreepada, "A Generic Frame Work for Image Data Clustering Via Weighted Clustering Ensemble," *International Journal of Computer Science & Information Technologies*, vol. 3, pp. 5429–5433, Nov. 2012.

[21] B. Pan and Z. Wang, "Recent Progress in Digital Image Correlation," in *Application of Imaging Techniques to Mechanics of Materials and Structures, Volume 4*, ser. Conference Proceedings of the Society for Experimental Mechanics Series. Springer New York, 2013, pp. 317–326.

[22] F. Hild and S. Roux, "Comparison of Local and Global Approaches to Digital Image Correlation," *Experimental Mechanics*, vol. 52, no. 9, pp. 1503–1519, 2012.

[23] R. Kloss, S. Silva, W. Schwartz, M. Cirne, and H. Pedrini, "Partial Least Squares Image Clustering," in *Conference on Graphics, Patterns and Images (XXVIII SIBGRAPI)*, Salvador-BA, Brazil, Aug. 2015.

[24] R. B. Kloss, M. V. M. Cirne, H. Pedrini, and W. R. Schwartz, "Low Redundancy and High Purity Image Clustering," *Pattern Recognition (under review)*, 2016.