Avgerinos, C., Nikolaidis, N., Mygdalis, V., & Pitas, I. (2016). Feature extraction and statistical analysis of videos for cinemetric applications. In Digital Media Industry & Academic Forum (DMIAF). Institute of Electrical and Electronics Engineers (IEEE). DOI: 10.1109/DMIAF.2016.7574926

Feature extraction and statistical analysis of videos for cinemetric applications

Christos Avgerinos†, Nikos Nikolaidis†, Vasileios Mygdalis† and Ioannis Pitas†⋆
† Department of Informatics, Aristotle University of Thessaloniki, Thessaloniki, 54124, Greece
⋆Department of Electical and Electronic Engineering, University of Bristol, UK

***Abstract*—In this paper, we describe a framework for the extraction of low-level and high level information from movies in order to be used for cinemetric applications. The developed framework analyses the available video content and extracts characteristics related to color, motion, contrast, shot length, tempo, face to frame ratios etc. The extracted information is stored in MPEG 7 AVDP profile format, which is a standard description format that can be imported to related cinemetric applications. We applied the developed framework in a collection of downloaded videos, as well as 3 stereoscopic movies.**

I. Introduction

Cinemetrics refers to the extraction of quantitative movie data characteristics. Cinemetrics can characterize movie segments (shots or frames) or collectively characterize an entire film. Derived data include low-level quantitative information (e.g., movie tempo, shot-length, shot type characterizations, face to frame ratio), or more high-level information (e.g., lead actor presence within a shot, shot activity information). The extracted characteristics can be subsequently used in various tasks, including movie genre classification, identification or recognition of specific director styles, movie-era recognition and analysis of movie style changes over time. For example, action films are expected to have higher movie tempo than drama films, and news videos are expected to have higher face to frame ratio than nature films [1]. Thus, these two low level characteristics can be utilized for movie genre classification.

Low-level features such as color histograms and motion features have been successfully used in video summarization and shot classification tasks [2]. Such features are also important in cinemetrics. Indeed, one of the key factors in describing visual content for cinemetrics is its color and motion information. More specifically, movie genres involve color palettes that distinguish them from the others. Color palettes describe the frequency of certain colors in a frame, a shot, or a movie. The mood of the audience is highly connected to the color characteristics of the scene, in terms of color hue, saturation or intensity [3]. Indeed, the viewer does not expect vivid oranges or greens when watching a crime or a horror movie neither does he expect many dark blues or blacks when laughing to a comedy film. The dominant color descriptor is used to define which colors are more frequently used in a scene [4]. On the other hand, movie genres differ with respect to the motion they include. For example, action or adventure movies are expected to feature rapid motion, while drama or romantic movies are not usually distinguished by such a trait [2]. In general, motion can be due to moving objects (or people) or movement of the camera [5], making extraction of dominant (object) motion cumbersome. The distinction of the dominant motion can become even more difficult if other moving elements such as text or graphics are included in the scene.

Nowadays, extracting meaningful low-level information from movies can be accomplished with significant success. For example, shot-boundary detection [6] is already considered solved or close to be solved. Moreover, combining information form shot cut and face detection algorithms can lead to successful shot-type classification (e.g., close up, long shot) [7].

In this paper, we describe a cinemetrics framework which extracts low-level quantitative information as well as high level information, in order to describe a movie. The derived semantic information is stored in MPEG 7 AVDP profile format [8], [9], which is an XML description format that can be easily plugged to any related cinemetric application. We evaluate the developed application in YouTube videos, as well as 3 stereoscopic movies.

In what follows, we present the information that can be extracted from the proposed framework in Section II. Example applications of the proposed framework can be found in Section III. Finally, conclusions are drawn in Section IV.

II. Information Extraction

We devised a framework for visual content description in terms of color, motion and other information, at the shot level. The developed color and motion descriptors, as well as the other employed descriptors are described in the following subsections.

A. Dominant Color

The dominant color descriptor we developed is based on HSV color space. Using HSV over RGB is a reasonable choice, as it is closer to the way that people perceive color information [10]. A low light scene, for example, is best described in HSV space, because the color and brightness are separated, which does not occur in RGB space. The dominant color descriptor is designed to be fast, thus, we employ quantization to a number of colors (e.g. 64) and resize the frame to 10%. Resizing the frame and keeping only a small percentage helps to limit the computations, and does not affect the results, as the color information is not highly affected by

image quality [5] [11]. Figure 1 present the original image and the quantized and resized copy.

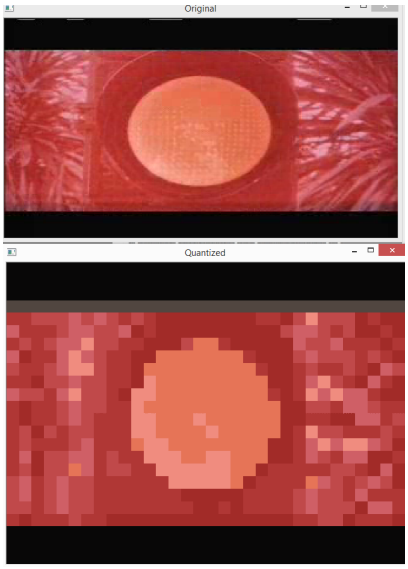


Fig. 1. Dominant color descriptor example

The dominant color descriptor works on a per-frame basis as follows. For each frame pixel, we use the hue, saturation and value triplet. The combination of hue, saturation and value with the highest frequency of appearance is the dominant color of the respective frame f^i,$H^{dom}i$, $S^{dom}i$, $V^{dom}i$. Apart from the dominant color, the descriptor also contains the mean and standard deviation values of each HSV channel separately. That is, within-frame information is also extracted with respect to the mean hue, saturation and value. Small standard deviation values denote that the frame colors are concentrated around the mean color.

B. Dominant Motion

The process of summarizing the motion activity of a number of frames or a whole movie, is based on optical flow. In the proposed framework, the optical flow is calculated by finding the moving pixels of the each frame and then calculating their new position on the next frame, as in [12]. Their movement is defined by the angle and the length of their trajectories. Observing every pixel of the frame would be time and resource consuming. By employing the Shi-Tomashi corner detector [13] we determine a number of strong corners of the image of a specific quality and distance. These variables were adjusted to distinguish true object motion from minor budges or slight camera motion. For the dominant motion, we employed the joint histogram of the angle and the length of the motion trajectories. The maximum value in this histogram expresses which movement is the most common between two frames.

Apart from the direction and the strength of the motion, we can also define the amount of activity in a film, with the visual disturbance value. The visual disturbance is the ratio of the moving pixels among the pixels that were selected from the Shi-Tomashi corner detector. Values close to 1 indicate movement of all the selected pixels, regardless of their strength or direction. An example of motion descriptor and its outputs can be seen in Figure 2.

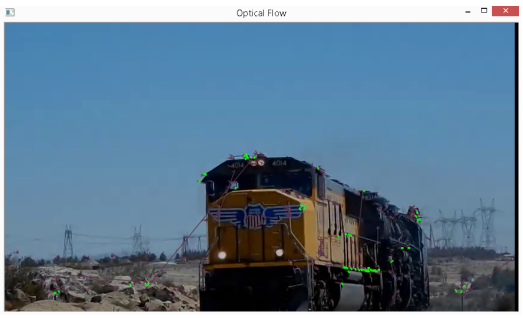


Fig. 2. Dominant motion example. The train is heading left at an angle of 153 degrees and a length of 2 units. This combination was found 55 times, which makes it the dominant motion.

C. Other extracted statistics

Apart from the color and motion features that were mentioned above, the proposed framework can extract a number of additional statistics. The basic cinemetrics statistic namely the *Shot Length* was evaluated by employing a shot detection algorithm [6] and evaluating the number of frames within each shot. Next, we implemented the *Rhythm* descriptor. The Rhythm is the number of shot changes in a predefined time interval. High rhythm values refer to movies with frequent shot changes, while low numbers indicate movies with slower tempo.

Moreover, by employing a modified version of the standard Viola-Jones face detector, that takes into account color information [14], we detected faces and subsequently evaluated the ratio of the largest face appearances in the duration of a shot to the frame area (*face to frame ratio*). In more detail, each frame of the shot is scanned for faces. Several faces may appear in a frame, so for every frame we only keep the largest one. In order to compute the size of a face, we use the facial bounding box (BB) area provided by the face detector, and subsequently evaluated the face to frame ratio. Having acquired information for each frame, it is relatively easy to translate it at the shot level, by employing the average face ratio from all the frames of the shot. In order to avoid having the shot results affected by a possibly very large (or a very small) BB, we first sort the vector that contains all the BB sizes for all the frames, and then select the median value, as in [1] [11]. Having extracted the face-frame ratios, we could use this information to extract the *shot-type* as in [7].

Finally, we extract the *contrast ratio*. In order to do so, we convert the input video to grayscale. Then, we quantize the grayscale values using a small number of bins, since using multiple levels would be time consuming and does not provide important information. Finally, the contrast ratio is evaluated as the difference of the min and max intensities in the frame. Similar to color information, contrast is a global descriptor of the image. Contrast information is important for movie genre

classification [15]. Horror or drama movies are expected to have a high contrast ratio because parts of the image are either abundant in light or very dark. On the other hand, comedy movies are expected to be characterized by low contrast values because the image brightness is more equally distributed in the frame.

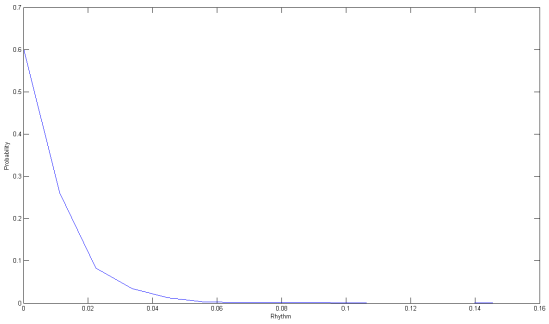
III. Statistical Analysis of Video Datasets

In order to apply the devised framework, we examined two possible scenarios. To this end, we have downloaded 1475 videos from YouTube, having random content. Moreover, we also employed three feature length stereoscopic movies. We extracted the cinemetric values with our framework in order to perform statistical analysis. In Subsections III-A and III-B and III-C, we present a rhythm, motion and face analysis of the YouTube videos, respectively. Finally, in Subsection III-D, we present a high level application scenario by employing the three movies, which were examined in terms of role identification.

A. Rhythm Analysis in YouTube videos

We computed the median rhythm value M_{rhythm} of all videos, being 0.004, meaning that the videos we used did not feature rapid shot changes. The rhythm average value μ_{rhythm} was found to be 0.008. The standard deviation σ_{rhythm} of rhythm values was found to be equal to 0.012, which indicates that rhythm values are close to the mean. Figure 3 shows the rhythm probability plot.

Fig. 3. Rhythm probability plot in YouTube videos

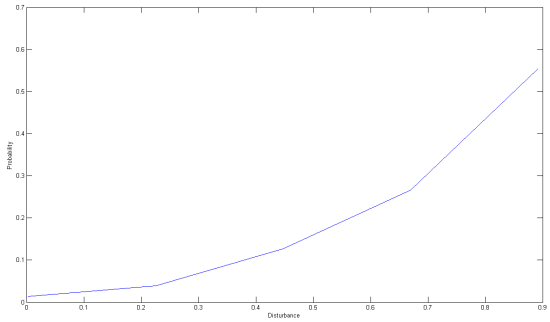


B. Motion analysis in YouTube videos

We split the motion analysis over the entire set of YouTube videos into two parts; the most dominant motion angle and the most dominant visual disturbance. The first value refers to the direction that the moving pixels usually follow, while the second is the most frequent number of moving pixels in a frame (out of those detected by the corner detector). For the visual disturbance analysis, we discarded the frames that contain no movement at all. Analysis showed that most involved movement is at an angle close to 0° (or 360°) meaning that most of the total motion is horizontal and to the right.

The respective most common visual disturbance value, dom_{dist} is at 0.8. This means that an average 80% of all the significant pixels did actually move, in the whole of our database. The median value M_{dist} was also computed at 0.8, while the mean value μ_{dist} was found to be 0.76. The standard deviation σ_{dist} was computed at 0.22, indicating that the visual disturbance values are concentrated close to the mean. Figure 4 shows the visual disturbance probability plot.

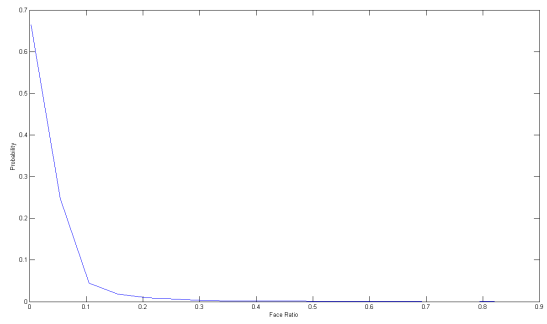
Fig. 4. Visual disturbance probability plot



C. Face to frame ratio statistics

For each of the YouTube videos, we extracted the median value of the face to frame ratio over all frames, leading to a value per video. Subsequently, we evaluated over the entire dataset, the most frequently appearing face to frame ratio value, which was found to be 2% (i.e., usually faces cover less than 2% of the frame area). The average face to frame ratio over all videos was 3.5%. Although no safe conclusions can be drawn due to the random nature of the videos in the YouTube dataset, these values indicate that persons depicted in these videos are usually captured in a long shot. Figure 5 depicts the face coverage probability plot.

Fig. 5. Face ratio probability plot



D. Role identification in movies

In this scenario, we have employed three feature length movies and examined the frequency of each actor appearance, i.e. the percentage of frames where this actor appears. All information was extracted from the proposed framework. Here

we should note that we applied a semi-automatic annotation procedure for each extracted facial image, giving actor labels by employing a label propagation algorithm [16]. That is, we employed a semi-supervised approach where we have manually annotated 10% of the exacted facial images and the remaining were automatically annotated. Then, we have split the movies in small video segments, where we reported the percentage of frames where the actor appeared in each segment. Finally for every actor in the movie we created a vector p that contains the appearance percentages. After normalization, we can assume that this one dimensional signal p can be viewed as a probability density function (pdf). Lead roles are expected to appear in almost every video segment of the movie, thus having a uniform distribution. Thus, we observe how much p deviates from the uniform distribution. To this end, we employ the Kullback Leibler divergence, KL, which is a measure of difference between two distributions. If a character appears on scattered parts of the movie, the KL divergence values between p and the uniform distribution should be low. On the other hand, appearances concentrated in some parts of the movie are expected to return high KL values.

Table I displays the KL values of the first 10 appearing actors in the employed movies. Here we should note that the employed movies belong to different genres, which are drama, action and adventure, respectively. As can be seen, the lead roles have similar low KL values in all 3 movies.

TABLE I
KULLBACK LEIBLER DIVERGENCE IN MOVIES

Actor	Movie1	Movie2	Movie3
Lead-Role	3.12	4.64	3.28
Co-star	12.09	5.45	12.16
Basic Role 1	20.85	24.57	20.20
Basic Role 2	26.89	28.33	20.40
Basic Role 3	30.61	33.59	21.00
Peripheral Role 1	31.36	36.58	21.90
Peripheral Role 2	35.90	37.37	26.17
Peripheral Role 3	37.39	38.9	32.01
Peripheral Role 4	38.99	40.43	32.28
Peripheral Role 5	40.45	40.46	33.64

IV. CONCLUSION

In this paper, we presented a framework which extracts low-level features as well as high level information from videos. The derived information can be used in a cinemetrics framework of other similar application areas. Future work could include expanding the number of supported statistics.

ACKNOWLEDGMENT

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement number 287674 (3DTVS)

REFERENCES

[1] X. Yuan, W. Lai, T. Mei, X.-S. Hua, X.-Q. Wu, and S. Li, "Automatic video genre categorization using hierarchical SVM," *IEEE International Conference on Image Processing (ICIP)*, pp. 2905–2908, 2006.

[2] D. Brezeale and D. J. Cook, "Automatic video classification: A survey of the literature," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 38, no. 3, pp. 416–430, 2008.

[3] C.-Y. Wei, N. Dimitrova, and S.-F. Chang, "Color-mood analysis of films based on syntactic and psychological models," *IEEE International Conference on Multimedia and Expo (ICME)*, vol. 2, pp. 831–834, 2004.

[4] B. Ionescu, K. Seyerlehner, C. Rasche, C. Vertan, and P. Lambert, "Content-based video description for automatic video genre categorization," 2012.

[5] T. Sikora, "The mpeg-7 visual standard for content description-an overview," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 11, no. 6, pp. 696–702, 2001.

[6] Z. Cernekova, I. Pitas, and C. Nikou, "Information theory-based shot cut/fade detection and video summarization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 16, no. 1, pp. 82–91, 2006.

[7] I. Tsingalis, A. Tefas, N. Nikolaidis, and I. Pitas, "Shot type characterization in 2d and 3d video content," *IEEE International Workshop on Multimedia Signal Processing (MMSP)*, 2014.

[8] I. Pitas, K. Papachristou, N. Nikolaidis, M. Liuni, L. Benaroya, G. Peeters, A. Roebel, A. Linnemann, M. Liu, and S. Gerke, "2d/3d audiovisual content analysis & description," *International Workshop on Multimedia Signal Processing (MMSP)*, pp. 1–6, 2014.

[9] K. Papachristou, N. Nikolaidis, I. Pitas, A. Linnemann, M. Liu, and S. Gerke, "Human-centered 2D/3D video content analysis and description," *International Conference on Electrical and Computer Engineering (ICECE)*, pp. 385–388, 2014.

[10] M. Montagnuolo and A. Messina, "Automatic genre classification of TV programmes using gaussian mixture models and neural networks," *International Workshop on Database and Expert Systems Applications*, pp. 99–103, 2007.

[11] G. Toderici, H. Aradhye, M. Paşca, L. Sbaiz, and J. Yagnik, "Finding meaning on youtube: Tag recommendation and category discovery," *Computer Vision and Pattern Recognition (CVPR)*, pp. 3447–3454, 2010.

[12] S. Vakkalanka, C. K. Mohan, R. Kumaraswamy, and B. Yegnanarayana, "Combining multiple evidence for video classification," *International Conference on Intelligent Sensing and Information Processing*, pp. 187–192, 2005.

[13] J. Shi and C. Tomasi, "Good features to track," *Computer Vision and Pattern Recognition (CVPR)*, pp. 593–600, 1994.

[14] G. Stamou, M. Krinidis, N. Nikolaidis, and I. Pitas, "A monocular system for automatic face detection and tracking," *Visual Communications and Image Processing 2005*, pp. 59 602C–59 602C, 2005.

[15] Z. Rasheed, Y. Sheikh, and M. Shah, "On the use of computable features for film classification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, no. 1, pp. 52–64, 2005.

[16] O. Zoidi, A. Tefas, N. Nikolaidis, and I. Pitas, "Person identity label propagation in stereo videos," *IEEE Transactions on Multimedia*, vol. 16, no. 5, pp. 1358–1368, 2014.