A Multi-Objective Optimization for Video Orchestration

Federico Colangelo, Federica Battisti, Marco Carli, Alessandro Neri
Department of Engineering
Università degli Studi Roma TRE
Roma, Italy
Email: {federico.colangelo, federica.battisti, marco.carli, alessandro.neri}@uniroma3.it

***Abstract*—In this work, the problem of video orchestration performed by combining information extracted by multiple video sequences is considered. The novelty of the proposed approach relies on the use of aesthetic features and of cinematographic composition rules for automatically aggregating the inputs from different cameras in a unique video. While prior methodologies have separately addressed the issues of aesthetic feature extraction from videos and video orchestration, in this work we exploit a set of features of a scene for automatically selecting the shots being characterized by the best aesthetic score. In order to evaluate the effectiveness of the proposed method, a preliminary subjective experiment has been carried out with experts from the audiovisual field. The achieved results are encouraging and show that there is space for improving the performances.**

***Index Terms*—Data analysis; multimodal signal processing; aesthetics**

I. Introduction

In the last four decades, social multimedia has been largely investigated since the rapid development of powerful, low-cost, consumer electronics products coupled with Internet, caused one of the largest social and technological transformations in society. New generations are experiencing interactive multimedia from an early age due the easy access to laptops, smart-phones, game consoles, and Internet connection. The social aspect is increasing importance and new systems are being developed for creating video sequences using shots from multiple camera sources. The goal of these systems is to automatically fuse footage of an event or of a scene taken from multiple cameras to create a richer social video experience.

The problem of sub-sequence orchestration is also faced by professional video-makers whose goal is to avoid missing important action or scene by exploiting events captured with multiple cameras. In literature, methods for combining shots from different cameras have been proposed. In [1], a supervised framework is presented in which the director selects the stream to be saved. In [2] an orchestration scheme for remote video-conferences is proposed. The system aims at automatically detecting the best camera from each of the locations in the conference. In [3], the authors propose an editing scheme based on low-sensor data as well as computer vision features to automatically edit video captured by mobile devices. In [4] the authors exploit editing rules concerned with continuity to perform video editing in the context of interacting videos.

In this work we propose an unsupervised system to perform an optimal orchestration based on aesthetic criteria and state-of-the-art rules adopted in cinema [5]. The goal of the generally adopted video editing guidelines is to produce a dynamic and smooth experience for the viewer. The aesthetic value of the scene is usually driven by the professional camera operator, following a shooting plan designed by the director. In our framework, the movie makers are not expert and the recording is not coordinated. Therefore, the video editing rules are coupled with an aesthetic evaluation of the scene.

As well known, the evaluation of the aesthetic is a very challenging task even for a human subject: beauty is subjective and there is not a well-defined set of rules according to which any image or video can be defined as *'beautiful'*. Furthermore, as shown in [6], aesthetic does not always correspond to the perceived quality: a video affected by severe blockiness will be addressed as low quality video, while a flat undistorted shot might have high quality but be aesthetically unappealing. In literature, this problem has been addressed exploiting machine learning based approaches. Datta et al. [7], adopt low-level features for predicting the aesthetic value of images, while a visual saliency map is introduced in [8]. In [9], the relative importance between foreground and background is considered. More recently, computational models of the aesthetic of consumer videos are designed in [6]. In [10], psycho-visual statistics extracted at different semantic levels are used as input for multiple classifiers. Finally, in [11], a systematic study to compare aesthetic assessments models, by taking into account the different experimental environments and rating scales is presented.

In this work we are dealing with a set of cameras recording the same scene from different points of view. The overall goal is to produce a video by exploiting the shots that in each time interval result more appealing from an aesthetic point of view while respecting the orchestration guidelines described in [5]. We propose to use a set of features for assessing the aesthetic value of each shot (e.g., a set of consecutive frames) recorded by each camera. This system can be useful for automatic video orchestration of amateur videos as well as a preprocessing step for media production.

The paper is structured as follows: in Section II the designed methodology is described along with the selected aesthetic feature. The validation for the proposed system is reported in

ISBN 978-0-9928626-7-1 © EURASIP 2017

Section III, while the results are presented and discussed in Section IV. Finally in Section V the conclusions are drawn.

II. Proposed Method

In the following, the term *scene* refers to the event being filmed. The term *shot* refers to the output of a single camera while *sub-shot* is used to indicate a part of the shot.

For all available video inputs, the position of each camera with respect to the recorded scene is described through the attributes listed in Table I. Since different camera may have a different acquisition setup (i.e., sampling rate or frame size), in the proposed system the K contributions are normalized with respect to the camera having the lowest frame rate f_s and the smallest frame size. The output video $\mathbf{V}$ is composed by N time-slots whose duration is adaptively determined based on the video content, as described in Subsection II-A; for each time-slot, the system selects one shot among the available K cameras. For reducing the computational complexity, N_f representative frames $\mathbf{R}_f$ are used for each second of the input videos.

While video editing is in general strongly dependent on the content and on the experience of the editor, there are guidelines in video editing theory [5], [12] for orchestrating multiple video sources. Usually, the main task of the editor is to create an entertaining and dynamic experience for the viewer. General rules derived from this principle are defined in [5] and can be summarized in the following:

- alternation of shot types and camera angles should be used to create dynamism;
- the length of a scene should be proportional to its content density. More dense scene take more time to become familiar (i.e. not entertaining) to the viewer;
- the angle of the camera should vary in at least 30 degrees between successive sub-shots;
- the angle should not vary too harshly as it creates confusion in the viewers.

Given these guidelines, we can articulate the orchestration problem in two parts: first, determine how often it is necessary to change video source (i.e. determine the duration of the time-slots); second, choose the content of each time-slot, maximizing the aesthetic and the compliance with the aforementioned rules.

A. Time-slot calculation

Basic rules of video editing state that changes of view should be used to create a more dynamic orchestration, therefore more changes are needed when the scene has a slow pace. The pace of the scene is estimated by extracting the Optical Flow (OF) from each of the available K videos.

The average OF for the K cameras is computed to obtain an estimation of the average dynamics of the content, $\overline{OF}$. This estimate is used for partitioning the video in three categories according to the motion rate: low, medium, and high dynamics as shown in Figure 1. The length of time slot is selected according to the $\overline{OF}$ value, as suggested in [5].

Shot Type	Very wide/Panoramic
	Wide
	Medium
	Close-up
	Point of view (POV)
Horizontal Angle (°)	0
	70
	140
	210
	280
Vertical Angle	Bird's eye
	High
	Eye-level
	Low
	Worm's eye

TABLE I: Camera framing features.

B. Time-slot content selection

To select the content of each time-slot, a multi-objective optimization is performed by using the Genetic Algorithm (GA). The algorithm operates over a set of possible candidates $\mathbf{V}_c$, evaluating the fitness through three functions:

- Mean aesthetic score of the sub-shots $A(\mathbf{V}_c)$
- Rule-based evaluation $R(\mathbf{V}_c)$
- Diversity score $D(\mathbf{V}_c)$

C. Aesthetic score computation

To assess the aesthetic value of a sub-shot, we use the approach proposed in [13] due to its low computational complexity. A subset of images from the CUHK dataset [14] is selected and, based on their labels, organized in two categories, high and low quality for creating a training set.

For each image, a 24-d feature vector is calculated and, based on the comparison with the high and low quality images in the reference dataset, the following information is calculated:

- Color palette (f_1): the color palette evaluates the color scheme in the HSV color space. A clustering is performed over the values of the color histogram and the calculated centroids are considered as the dominant colors. f_1 is obtained by comparing dominant colors in the image under analysis with the ones extracted from the training set.
- Layout Composition ($f_2 - f_5$): a layout template is extracted from the available high and low quality images by averaging the value of the image coefficients over four channels (H,S,V and H+S+V). The features are obtained by calculating the L_1 distance of the image under analysis, d_H and d_L respectively, from the templates.
- Edge Composition ($f_6 - f_9$): the edge features are obtained by averaging the edge information of the high and low quality images to obtain reference templates. Image features are extracted by subtracting the L_1 distances from the templates.
- Global texture ($f_{10} - f_{17}$): these features are generated by dividing the image into 6 stripes and computing the sum of the differences of adjacent stripes.
- General features ($f_{18} - f_{24}$): features evaluating the amount of blur, contrast, and the number of non-zero

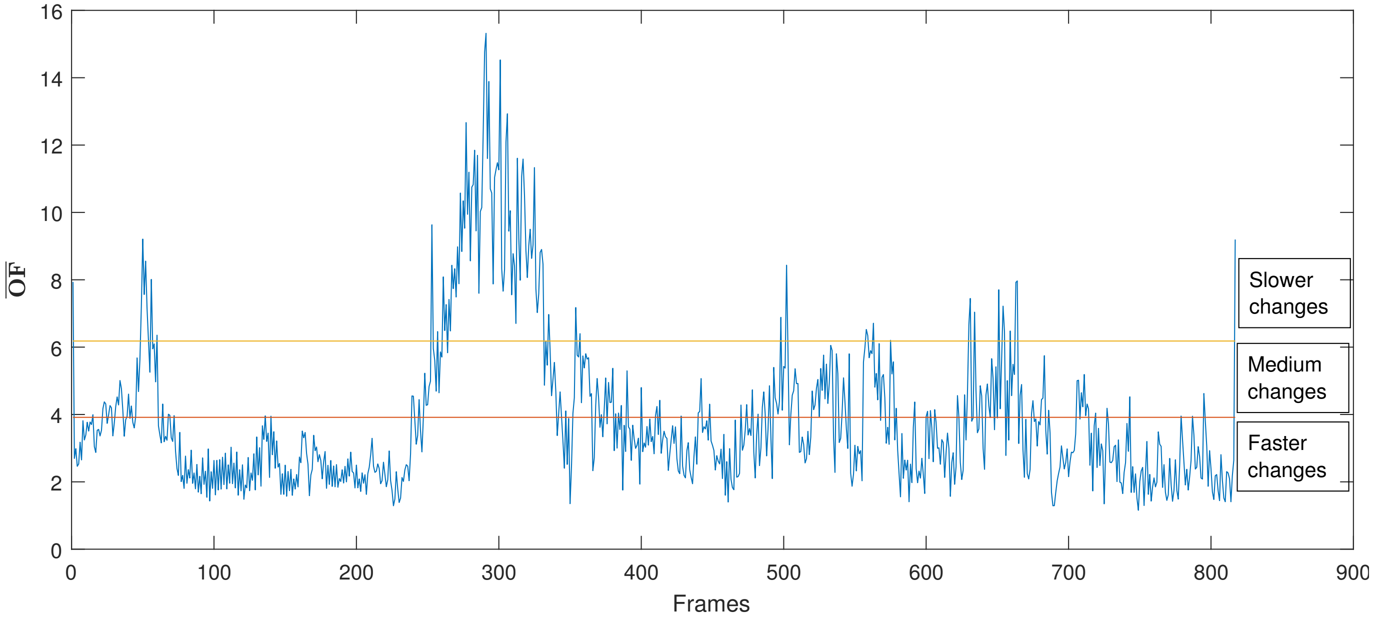

Fig. 1: Motion rate partitioning

elements in the HSV quantized histogram and the dark channels [15].

The resulting feature vector is classified with a SVM (Support Vector Machine). In our work, the aesthetic score of a sub-shot is evaluated by classifying the features extracted from the $\mathbf{R}_f$ frames in probabilistic mode, assigning a default label of high quality. The classifier returns the predicted class together with the confidence level, expressed as a probability. The confidence level is used as aesthetic score.

D. Rule-based evaluation

The purpose of $\mathbf{R}(V_c)$ is to evaluate the smoothness of the transitions in V_c. In order to do this, we model V_c as a Markov chain over the possible framings of the scene. Each state (i.e. framing) is characterized by the attributes used to tag the cameras in Table I. The distance between two states is given by the sum of the differences in the values of the considered attributes. Nearby states will be cameras with similar features. Transitions to nearby states have a high probability, while lower values are given to far states (i.e. very different framing). Based on the cinematographic rules, we impose that the probability of remaining in the same state is low ($< 10^{-4}$). The distance between states is determined adaptively depending on which type of camera attributes are given as input: a panoramic shot and a POV shot could be considered near only if no other intermediate shot type (e.g. medium shot) is given. In this way, the system prefers smooth transitions in the view, avoiding harsh scene changes that would be annoying for the viewer. The value of $R(\mathbf{V}_c)$ is given by the probability of the correspondent path ni the Markov chain:

$$R(\mathbf{V}_c) = \sum_{i=1}^{N-1} \log(P(\mathbf{V}_c(i), \mathbf{V}_c(i+1)))$$

where P is the probability of transition from $\mathbf{V}_c(i)$ to $\mathbf{V}_c(i+1)$.

E. Diversity evaluation

In a real scenario, it can happen that one of the videos has an overall higher aesthetic score with respect to the others. This may be due to several reasons such as one of the operator being more skilled or be in a better position with respect to the others. In this case, performing an optimization based exclusively on $A(V_c)$ and $R(V_c)$ would result in the exclusion of a consistent number of video sources.

Preliminary tests carried out with 10 experts, have shown that a video obtained as orchestration of diverse inputs is preferable to one composed by combining a fewer number of cameras, even if they are more valuable in aesthetic. For this reason we introduce a third function in the optimization problem, the diversity score, to take into account and penalize the unbalanced use of cameras. The diversity score $D(\mathbf{V}_c)$ is calculated through the following steps: first the cameras empirical probability distribution $P_{\mathbf{V}_c}$ is calculated. $P_{\mathbf{V}_c}$ is then compared with a uniform probability distribution, P_u, over the K cameras through the Kullback-Leibler (KL) divergence:

$$D_{KL}(P_u, P_{\mathbf{V}_c}) = \sum\nolimits_{i \in K} \ln\left(\frac{P_u(i)}{P_{\mathbf{V}_c}(i)}\right) P_u(i).$$

F. Multi-objective optimization

The goal of finding an optimal $\mathbf{V}$ can be formulated as a multi-objective optimization problem over the three previously defined functions, $A(\mathbf{V}_c)$, $R(\mathbf{V}_c)$, $D(\mathbf{V}_c)$. This class of problem does not have a unique optimal solution. In fact, the solutions produced by a multi-objective optimization are Pareto-optimal, that is a solution where none of the involved functions can be optimized without degrading the others. Evolutionary algorithms are often used in this setup since they can find multiple

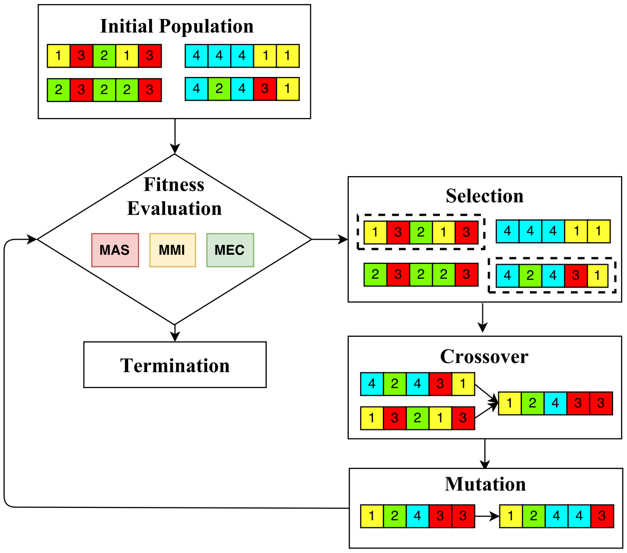

Fig. 2: Multi-objective genetic optimization of the editing

Video set id	Cameras	$f_s(fps)$	Length (s)	Res. (pxl)
1	4	29	6	1080x1080
2	3	29	10	1280x720
3	4	25	10	1080x1080
4	4	25	10	1080x1080

TABLE II: Source videos used for the experiment

non-dominated solutions with each iteration [16]. To find a set of candidate vectors $\mathbf{V}_c$, we perform a multi-objective optimization using the GA.

The GA is inspired to natural selection processes, taking a populations of individuals with different sets of genes. In our case, the population is composed by candidate editing vectors $\mathbf{V}_c$, where a gene is the sub-shot contained in a time-slot. A fitness value is then calculated for each individual in the population, in our case using the $A(\mathbf{V}_c)$, $R(\mathbf{V}_c)$, $D(\mathbf{V}_c)$ functions.

Individuals with higher fitness scores are more likely to be used as *parents* in the crossover step. The next generation is calculated during crossover by combining the genes of the parents. The final step of an iteration is the mutation: genes can be changed to random values with a small chance. The optimization problem described above can be formalized as:

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & -A(\mathbf{V}_c), -R(\mathbf{V}_c), D(\mathbf{V}_c)) \\ \text{subject to} \quad & 1 \leq \mathbf{V}_c(i) \leq K, \; i = 1, \ldots, N \end{aligned}$$

this approach can generate multiple locally optimal solutions. We consider this feature desirable, as multiple possible editing can be examined by professionals, fine tuning the orchestration to a desired output.

G. Time-slot duration fine tuning

The duration of each time-slot is determined without knowing its content. Nevertheless, after the optimization procedure, the duration of time-slots can be fine tuned based on the content of $\mathbf{V}$. In more details, longer time-slots are assigned to cameras that are richer in details, since editing guidelines state that denser scenes take longer to become boring for the viewer.

In order to estimate the content density of a scene, we extract the mean edge values E_d in the first frame of each sub-shot:

$$E_d(\mathbf{V}(i)) = \frac{1}{m*n} \sum_{i=1}^{n} \sum_{j=1}^{m} \nabla^2 f_{\mathbf{V}(i)}(1, i, j).$$

Where $f_{\mathbf{V}(i)}(1)$ is the first frame of $\mathbf{V}(i)$ and m, n are the width and height of the frame. For each pair of consecutive sub-shots $V_c(i)$, $V_c(i+1)$ the sub-shot denser in content has its time-slot length, t_d extended by Δt. Nevertheless, this adjustment is performed only if the shortened sub-shot length is still longer or equal to 1 second to avoid introducing too short sub-shots.

III. System Validation

In order to verify the effectiveness of the proposed method, a preliminary test has been performed. Four different contents were considered and for each content different views were taken. Table II gives specifications for the video sets used to test the algorithm. In this trial, subjective tests were performed with the cooperation of 16 experts in media production and movie generation. As described in [17] even if the results of expert viewing cannot be considered as a replacement of the results provided by a formal subjective assessment, they can be considered a valuable preliminary indication of the performances of the systems under test. Due to the complexity of the problem of subjective evaluation of aesthetic in videos, we decided to run an expert viewing test before running a full subjective test. All the subjects are between 25 and 35 years old, 8 of the subjects were male and 8 were female. The video set 1 portrays a roller coaster ride recorded from four different points of view, selected according to cinematographic rules: subjective point of view, action of the subject, details of the subject and panoramic view. The video set 2 contains the final moments of a concert, with three views: left side and close to the stage, left side and far from the stage and central and close to the stage. The video set 3 records from four points of view the scene in a cafeteria in which a guy is drinking a coffee. Finally, the video set 4 plays a scene in which a bartender opens a bottle of wine. The video sets 1, 3, and 4 were recorded by using commercial mobile phones and a Go-Pro camera while the video set 2 was gathered from the internet. From each video content, by considering all the available views, two different orchestrations were performed: one based on the proposed algorithm and another one obtained by randomly merging the input views. This means that overall eight contents were used in the subjective experiment. For each video content, the two orchestrated videos were shown to each user that was asked to selected the one that, according to his/her experience was preferable. Furthermore an interview was performed for collecting the feedback. The display used

ISBN 978-0-9928626-7-1 © EURASIP 2017

Video set id	Proposed (%)	Random (%)
1	68.8	31.2
2	56.3	43.7
3	18.8	81.2
4	25	75

TABLE III: Percentage of preferences expressed during the subjective tests.

for the experiment was a Full HD display. The parameters used in the experiment are: $N_f = 4$, faster, medium and slower change rates are set to 1s, 2s, and 3s respectively. For the GA, an initial population of 15 individuals has been used.

IV. Results and discussion

From the collected results, reported in Table III, an important feedback can be extracted. It is possible to highlight two different trends. For the video sets 3 and 4 the randomly orchestrated videos are preferred. This is mainly due to the fact that the system was optimized for dealing with dynamic content characterized by fast scene changes, while the video sets 3 and 4 are characterized by slow motion. It results that a preliminary analysis of the video content and of the motion rate should be performed in order to optimize the orchestrator performances, and to adapt the scene changing rate to the video dynamics. This behavior is confirmed by the results obtained for the video sets 1 and 2. In this case, the content is dynamically changing and the videos generated by the proposed orchestrator are preferred. Information about content also allows a semantic, template-based approach, similar to the one used in [13], for video-based aesthetic feature. As an example, having a reference template of motion for action videos would allow to use an aesthetic fitness function leveraging also video information.

V. Conclusions

In this contribution, an automatic system for orchestrating video inputs taken from different devices and different points of view is presented. It relies on the use of aesthetic features for selecting the best shot among the available ones and of cinematographic composition rules.

In order to evaluate the effectiveness of the proposed method, a preliminary subjective experiment has been carried out with experts from the audiovisual field. The achieved results show that the system is strongly dependent on the dynamics of the input videos. For this reason, a way forward for improving the system is to adapt the scene changing rate to the video dynamics.

References

[1] B. Gong, W. Chao, K. Grauman, and F. Sha, "Diverse sequential subset selection for supervised video summarization," in *Proceedings of the Neural Information Processing Systems (NIPS)*, 2014.

[2] R. Kaiser, P. Torres, and M. Höffernig, "The interaction ontology: Low-level cue processing in real-time group conversations," in *2nd ACM International Workshop on Events in Multimedia*. EiMM '10, ACM.

[3] W. Taylor and F. Z. Qureshi, "Automatic video editing for sensor-rich videos," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, March 2016.

[4] E. S. d. Lima, B. Feij, A. L. Furtado, A. Ciarlini, and C. Pozzer, "Automatic video editing for video-based interactive storytelling," in *2012 IEEE International Conference on Multimedia and Expo*, July 2012, pp. 806–811.

[5] K. Dancyger, *The Technique of Film and Video Editing History, Theory, and Practice.*

[6] A. K. Moorthy, P. Obrador, and N. Oliver, "Towards computational models of the visual aesthetic appeal of consumer videos," in *Proc. of Computer Vision, ECCV 2010.*

[7] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Studying aesthetics in photographic images using a computational approach," in *Computer Vision ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006.*

[8] Z. Dai and Y. Wu, "Where are focused places of a photo?," in *Advances in Visual Information Systems: 9th International Conference, VISUAL 2007 Shanghai, China, June 28-29.*

[9] P. Obrador, "Region based image appeal metric for consumer photos," in *Proc. of Multimedia Signal Processing, 2008 IEEE 10th Workshop on.*

[10] S. Bhattacharya, B. Nojavanasghari, T. Chen, D. Liu, S.-F. Chang, and M. Shah, "Towards a comprehensive computational model foraesthetic assessment of videos," in *Proc. of the* 21^{st} *ACM International Conference on Multimedia*, 2013.

[11] E. Siahaan, A. Hanjalic, and J. Redi, "A reliable methodology to collect ground truth data of image aesthetic appeal," *IEEE Transactions on Multimedia*, vol. 18, no. 7, pp. 1338–1350, July 2016.

[12] W. Murch, *In the Blink of an Eye: A Perspective on Film Editing.*

[13] K.-Y. Lo, K.-H. Liu, and C.-S. Chen, "Assessment of photo aesthetics with efficiency," in *Pattern Recognition (ICPR), 2012 21st Int. Conf. on.*

[14] Wei Luo, Xiaogang Wang, and Xiaoou Tang, "Content-based photo quality assessment," in *Computer Vision (ICCV), 2011 IEEE Int. Conf. on.*

[15] X. Tang, W. Luo, and X. Wang, "Content-based photo quality assessment," *IEEE Transactions on Multimedia*, vol. 15, no. 8, pp. 1930–1943, 2013.

[16] D. Kalyanmoy, *Multi-Objective Optimization Using Evolutionary Algorithms*, John Wiley & Sons, Inc., 2001.

[17] ITU, "ITU-R Recommendation, Subjective assessment of video quality using expert viewing protocol," BT.2095, Apr. 2016.

ISBN 978-0-9928626-7-1 © EURASIP 2017