RECfusion: Automatic Video Curation Driven by Visual Content Popularity

Alessandro Ortis
Universitá degli Studi di Catania
Viale A. Doria 6, 95125
Catania, Italy
ortis@dmi.unict.it

Giovanni Maria Farinella
Universitá degli Studi di Catania
Viale A. Doria 6, 95125
Catania, Italy
gfarinella@dmi.unict.it

Valeria D'Amico
Telecom Italia
JOL WAVE - Viale A. Doria 6, 95125
Catania, Italy
valeria1.damico@telecomitalia.it

Luca Addesso
Telecom Italia
JOL WAVE - Viale A. Doria 6, 95125
Catania, Italy
luca.addesso@telecomitalia.it

Giovanni Torrisi
Telecom Italia
JOL WAVE - Viale A. Doria 6, 95125
Catania, Italy
giovanni.torrisi@telecomitalia.it

Sebastiano Battiato
Universitá degli Studi di Catania
Viale A. Doria 6, 95125
Catania, Italy
battiato@dmi.unict.it

ABSTRACT

The proliferation of mobile devices and the diffusion of social media have changed the communication paradigm of people that share multimedia data by allowing new interaction models (e.g., social networks). In social events (e.g., concerts), the automatic video understanding goal includes the interpretation of which visual contents are the most popular. The popularity of a visual content depends on how many people are looking at that scene, and therefore it could be obtained through the "visual consensus" among multiple video streams acquired by the different users devices. In this work we present RECfusion, a system able to automatically create a single video from multiple video sources by taking into account the popularity of the acquired scenes. The frames composing the final popular video are selected from the different video streams by considering those visual scenes which are pointed and recorded by the highest number of users' devices. Results on two benchmark datasets confirm the effectiveness of the proposed system.

Categories and Subject Descriptors

I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding—*Perceptual reasoning, Video Analysis*; I.4.9 [**Image Processing And Computer Vision**]: Applications

Keywords

Social Cameras, Video Curation, Summarization

MM'15, October 26–30, 2015, Brisbane, Australia.

DOI: http://dx.doi.org/10.1145/2733373.2806311.

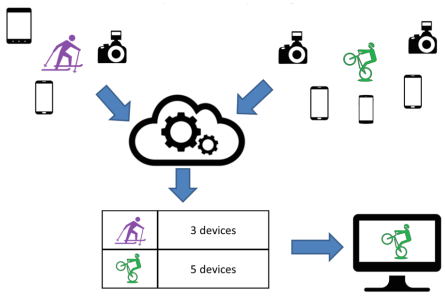

Figure 1: RECfusion - Automatic summarization of popular scenes of social events in a single video.

1. INTRODUCTION

In social events like a concert, a maratona or a car ride, the audience gathers multimedia information with mobile devices (e.g., images, video, geolocation, tags, etc.) related to what has captured their interest. Popular scenes (e.g., fireworks in a folkloristic event) are often observed and acquired simultaneously by multiple end-users with different devices. This redundancy in the video sequences can be exploited to infer which groups of people are interested to specific visual contents over time, and hence which scenes are the most popular.

This work proposes a system that automatically processes multiple video flows from different devices to understand the most popular scenes for a group of end-users. The scenes observed by the different devices are grouped by visual content. This allows an automatic video curation process to obtain a single video as output, by mixing the different inputs and taking into account the most popular scenes, i.e. those scenes acquired by many devices over time (Figure 1). Although much effort has been devoted to understand what is interesting in a scene [7], this problem is still unsolved. The task of establishing the popularity of a scene is challenging because of the variability of the visual content observed by multiple devices: different points of view, pose and scale of the objects, lighting conditions, occlusions, viewing quality, as well as different device models. The Imaging Generation Pipeline (IGP) can vary from device to device and even on

an per-image basis [5]. The difference between responses depends essentially on the characteristics of the lenses, filters (e.g. Bayer pattern), sensors, and IGP algorithms [6].

The video curation system proposed in this paper works without specific priors on the input, as well as without knowledge of which and how many devices are involved in the curation process. A related work is described in [1], where a 3D reconstruction of the scene and the relative pose of the devices are exploited. However, the reconstruction of the whole scene is computationally expensive and the video curation algorithm in [1] can be applied only after such a preprocessing stage. In [10] it is proposed an approach which exploits multiple devices of the same model. Also in this case the algorithm needs a calibration phase to reconstruct the camera poses. The method proposed in [8] tries to create a single popular video of an event from multiple egocentric videos. All scenes were taken using the same camera model by different participants close each other (few meters). The approach assumes that the number of the different popular scenes and the number of the devices is known a priori. These information are used to recognize an exact number of regions of interest used as prototypes in the grouping phase. The algorithm proposed in [11] combines several videos of the same scene taken from different perspectives. Despite the final video is not based on the popularity (as in the case we are interested in) this framework is able to produce an unique final video related to the considered scenario. The main goal of the approach in [11] is to build an automatic system which tries to imitate a professional video editor by choosing automatically which shooting angle and distance should be used and how long the selected configuration should persists.

Differently than previous approaches, the method proposed in this paper combines several videos from unknown different devices based on the popularity of the acquired scenes without any prior knowledge or training stage.

2. PROPOSED SYSTEM

The multiple video streams acquired by the different devices are analysed by using two algorithms: one to segment the different scenes, transitions and the unstable intervals within each video (intraflow analysis), and the other one to perform the grouping of the involved devices over time by taking into account the visual content of the previously segmented video streams. The popularity of the obtained clusters over time is used to produce the final video.

2.1 Intraflow Analysis

During intraflow analysis each video is processed by comparing its frames in order to segment the video based on the visual content. For each frame selected by sampling the video, we extract keypoints using the well-known SIFT algorithm [9]. The set of SIFT features extracted from a frame are used as a template for the acquired scene. In our experiments we excluded the SIFT extracted near the border of the considered frames to make more robust the feature matching among frames. The intraflow analysis consists in the comparison of templates extracted from different frames of a video to split it in blocks by taking into account the visual content. During this process the system keeps a reference template regarding the last known scene (i.e., the last stable set of SIFT features extracted from the last detected scene) and compares this template with respect to the features extracted from the current frame under analysis. When a sensible variation of features is observed (i.e., low matching score), the algorithm refreshes the reference template and splits the video producing a new segment.

Given a frame, the number of matchings between its SIFT keypoints and the once of the reference template is considered as a similarity index between the involved scenes. To make the matching more reliable, we excluded the matchings where the keypoints are too far in terms of spatial coordinate by assuming smooth transition between frames (we used a threshold distance of 100 pixels for images with resolution 1280×720 or 1920×1080). In order to detect the sudden changes of the number of matchings we defined a slope function which is computed on a frame at time T as follows:

$$slope(T) = \frac{h}{w} = \frac{l \sin\theta}{l \sin\theta} = \tan\theta \quad (1)$$

This function represents the variation of the number of matchings h in a range interval w centered in a frame at time T. This value is related to the tangent of an angle θ which is proportional to the gradient of the matching curve. The algorithm asks for a new template (i.e., set of SIFT features) when the slope function has a peak greater than 10 (i.e., $\theta = 85°$). In order to define only reliable templates the algorithm checks if the computed template is stable for at least 2 seconds (i.e., the number of matchings do not change too much). When a new stable template is defined, the algorithm compares it with respect to the past templates in order to understand if it regards a new scene or it is related to a known previously observed scene. This backward checking is done starting from the last found template. Two different templates of the same scene could be rather different due to the elapsed time between them. During this step, to check if two templates describe the same scene we use a geometric verification to exclude the spatial matchings with distance higher than one-third of the height of the image. Two templates are assigned to the same scene if the percentage of the matchings after the geometric verification is greater than 50% of the original matchings. Each reference template is assigned to a scene ID and all the video frames which achieve a robust match with a reference template are classified as of that scene. All the frames between the instant when a new scene template have to be upgraded by the system and the instant when that template is finally upgraded are classified as a transition interval.

Figure 2 shows an example of the result of the intraflow analysis applied on four input videos as a coloured chronogram: each scene is identified by a colour (red, blue and green in the figure), whereas the transition intervals and the unstable frames (e.g., shaking frames) are identified by black colour. The intraflow analysis allows an automatic segmentation of each video in several intervals depending on the visual content. It is also useful to correctly locate the transition intervals.

2.2 Interflow Analysis

Given two images acquired with different devices it is very challenging to understand if they are related to the same scene using only visual information [3, 4]. In our case, given frames of different videos which have been segmented as described in Section 2.1 we want to understand which of the different devices are looking at the same scene over time. The most popular scene over time is then used to produce the final video. In the interflow analysis we defined a frame descriptor based on a weighted colour histogram. In order

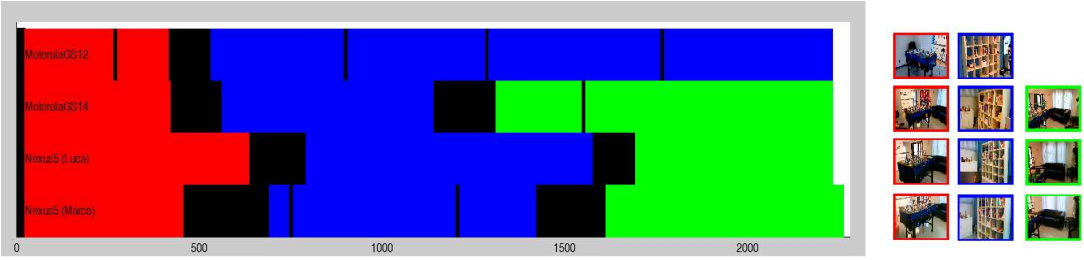

Figure 2: Output of the intraflow analysis.

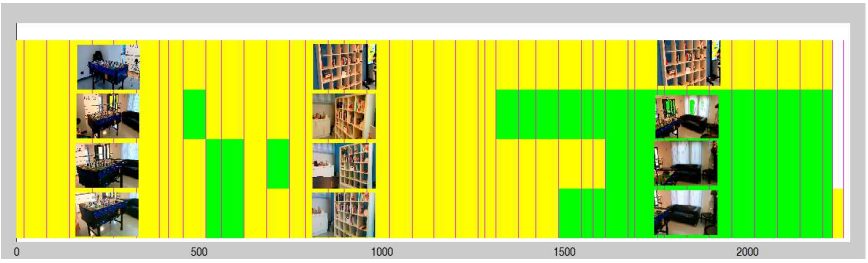

Figure 3: Clustering result applied on four devices. A vertical red line indicates when a clustering block is performed over time, this happens when one of the devices change his subject (e.g. in corrispondence of a intraflow change in Figure 2) or at most every 2 seconds.

to address the device invariance issue, we first apply the histogram equalization described in [5]. This method consists on the application of an histogram equalization to each RGB channel to reduce the variability introduced by the IGPs related to the different devices. After the equalization we compute a weighted colour histogram by quantizing the color space (8 color for each channel). The weights are obtained by using a gradient map as suggested in [2]. The gradient map highlights the structures of the objects involved in the scene making more robust the color-based descriptor. To compare histograms we use the same distance employed in [2]:

$$d(h_{D_a}, h_{D_b}) = \sum \frac{\sum \left(h_{D_a} - h_{D_b}\right)^2}{\sum \left(h_{D_a}\right)^2} \qquad (2)$$

where $h_{\mathrm{D_a}}$ and $h_{\mathrm{D_b}}$ are the weighted histograms related to the two frames of two different devices D_{a} and D_{b}.

To cluster the devices accordingly to visual content at every instance of time, we first segment all the videos exploiting the intraflow analysis (Section 2.1), and then we use the weighted color histogram representation to compare the obtained video segments (Figure 3). Through the intraflow analysis we obtain the scene ID and a set of SIFT features used as scene template for each frame of the different videos. At each instant T, the set of frames of the different devices are clustered by taking into account the corresponding weighted histogram representation. The different scenes are considered as a complete graph where each node is a device and the arches are labelled with the interflow measure between the scenes taken by the devices (Equation 2). The interflow measures are used to establish similarity during the clustering of the devices. The final video produced as output by the proposed approach is obtained by considering the most popular cluster (i.e., the once with the highest number of devices). Specifically, for the output at each instant T we consider the video belonging to the most popular cluster which is closest to the cluster centroid (i.e., average among all histograms). The pseudocode to cluster the devices at each instant of time is reported in Algorithm 1.

Algorithm 1 Devices' clustering

```
procedure
    Choose an unclustered device D.
    Insert D in a new cluster C_d.
    for each device B ≠ D do
        if d(h_D,h_B)<1 then
            if B is unclustered then
                Insert B in the cluster C_d.
            else
                Find the cluster containing B.
                Compare B with the elements
                of the two contending clusters.
                Insert B in the closer cluster.
        if all devices are clustered then
            End procedure.
        else
            Return to 2.
```

3. EXPERIMENTAL RESULTS

To demonstrate the effectiveness of the proposed system we have performed experiments on a new benchmark dataset available at the URL http://iplab.dmi.unict.it/recfusion. The considered dataset includes different scenarios acquired with multiple devices of different models:

- Foosball - In this scenario each device takes a view of a Foosball room switching among three different regions of interest. This scenario is useful to highlights the behaviour of the system when the popularity of a scene leaves a subject advocating a new one.
- Meeting - There were four people sitting around a table. Each person records one of the participants using a mobile device (smartphone or tablet). One more device is placed in order to look constantly to a subject. We tested this scenario varying the number of the devices (2, 4 or 5 devices).
- SAgata - This set of videos have been taken in a real scenario during a folkloristic event (Saint Agata in Catania). Seven people recorded the event while the main subject (the Agata's statue) was carried along the city's streets.

Figure 4: Output video example. On the left and right the input videos, the final output at the center.

The aforementioned scenarios include challenging scenes with crowd, large perspective variation, occlusion, periods of tilt and shaking, etc. We have also tested the proposed approach on the benchmark dataset proposed in [8].

For testing purposes each video has been manually segmented. For each instant of time we known the exact number of clusters and the most popular scene. To evaluate the performances of the proposed method on each scenario we computed two measures obtained from the following scores calculated on each clustering over time:

- P_{r}: ground truth popularity score (number of cameras looking at the most popular scene) obtained from manual labelling;
- P_{a}: popularity score computed by our method (number of the elements in the popular cluster);
- P_{g}: number of correct videos in the popular cluster computed by our method.

From the above scores, we computed the weighted mean of the ratios $\mathrm{P_a/P_r}$ and $\mathrm{P_g/P_r}$ over all the segmented blocks of a video, where the weights are given by the length of the blocks (i.e., the number of frames). When $\mathrm{P_a/P_r}$ is close to 1, the popularity score computed by our method is similar to the ground truth popularity. When this number is greater than 1 it means that the most popular cluster obtained with our approach is affected by outliers, whereas when this number is less than 1 it means that our method missed some element of the ground truth popular cluster. Since $\mathrm{P_a/P_r}$ deal just with the number of video in the popular cluster, it is useful to look also at the ratio $\mathrm{P_g/P_r}$. Indeed, $\mathrm{P_g/P_r}$ assesses the visual content of the videos in the popular cluster (i.e., true positive). This score have to be close to 1 to indicate accuracy in the popular cluster computed by our method.

Table 1 shows the obtained results. The first five rows are related to the scenarios of the proposed dataset, whereas the last three rows are related to the dataset proposed in [8]. The results show the effectiveness of our approach. Difficulties appear when some video regarding the most popular subject are taken with a quite different scale factors. This can be noted comparing the third and the fourth row in Table 1. In the meeting scenario (5 devices of different models) there is a huge difference in the scale of the acquired subjects in the scene. In the videos proposed in [8] the camera is constantly moving due to the shake induced by the natural head motion of the wearer. Despite we achieve good performances on wearable egocentric videos, we believe that there is still space for further improvements in such a video category (e.g., by filtering out the head motion).

In order to better asses the results obtained by the proposed system, the reader can perform a visual inspection of the videos produced by our approach at the following URL: http://iplab.dmi.unict.it/recfusion.

Table 1: Experimental Results

Scenario	Devices	Models	$\mathrm{P_a/P_r}$	$\mathrm{P_g/P_r}$
Foosball	4	2	1.02	1
Meeting	2	2	1.01	0.99
Meeting	4	4	0.99	0.95
Meeting	5	5	0.89	0.76
SAgata	7	6	1.05	1
Concert [8]	3	1	1.06	1
Lecture [8]	3	1	1.05	0.86
Seminar [8]	3	1	0.62	0.62

4. CONCLUSIONS

This work proposed an automatic video curation method driven by the popularity of the scenes acquired by multiple devices. Although some errors could occur during the clustering of the devices, the system rarely chooses outlier video frames as the output for the proposed dataset. When the algorithm works with a few number of input video these errors could affect the popularity of the clusters. However, if the number of the devices increases, then the effect of the clustering errors is reduced.

5. ACKNOWLEDGMENTS

This work has been performed in collaboration with Telecom Italia JOL WAVE in the project FIR2014-UNICT-DFA17D.

6. REFERENCES

[1] I. Arev, H. S. Park, Y. Sheikh, J. Hodgins, and A. Shamir. Automatic editing of footage from multiple social cameras. *ACM Transactions on Graphics*, 33(4):81, 2014.

[2] J. Domke and Y. Aloimonos. Deformation and viewpoint invariant color histograms. In *British Machine Vision Conference*, 2006.

[3] G. M. Farinella and S. Battiato. Scene classification in compressed and constrained domain. *IET Computer Vision*, 5(5):320–334, 2011.

[4] G. M. Farinella, D. Ravì, V. Tomaselli, M. Guarnera, and S. Battiato. Representing scenes for real-time context classification on mobile devices. *Pattern Recognition*, 48(4):1086–1100, 2015.

[5] G. Finlayson, S. Hordley, G. Schaefer, and G. Y. Tian. Illuminant and device invariant colour using histogram equalisation. *Pattern recognition*, 38(2):179–190, 2005.

[6] G. Finlayson and G. Schaefer. Colour indexing across devices and viewing conditions. In *Workshop on Content-Based Multimedia Indexing*, 2001.

[7] H. Grabner, F. Nater, M. Druey, and L. Van Gool. Visual interestingness in image sequences. In *ACM International Conference on Multimedia*, 2013.

[8] Y. Hoshen, G. Ben-Artzi, and S. Peleg. Wisdom of the crowd in egocentric video curation. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2014.

[9] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[10] H. S. Park, E. Jain, and Y. Sheikh. 3d social saliency from head-mounted cameras. In *Advances in Neural Information Processing Systems*, 2012.

[11] M. K. Saini, R. Gadde, S. Yan, and W. T. Ooi. Movimash: online mobile video mashup. In *ACM International Conference on Multimedia*, 2012.