TUD-MMC at MediaEval 2016: Predicting Media Interestingness Task

Cynthia C. S. Liem
Multimedia Computing Group, Delft University of Technology
Delft, The Netherlands
c.c.s.liem@tudelft.nl

ABSTRACT

This working notes paper describes the TUD-MMC entry to the MediaEval 2016 Predicting Media Interestingness Task. Noting that the nature of movie trailer shots is different from that of preceding tasks on image and video interestingness, we propose two baseline heuristic approaches based on the clear occurrence of people. MAP scores obtained on the development set and test set suggest that our approaches cover a limited but non-marginal subset of the interestingness spectrum. Most strikingly, our obtained scores on the Image and Video Subtasks are comparable or better than those obtained when evaluating the ground truth annotations of the Image Subtask against the Video Subtask and vice versa.

1. INTRODUCTION

The MediaEval 2016 Predicting Media Interestingness Task [3] considers interestingness of shots and frames in Hollywood-like trailer videos. The intended use case for this task would be to automatically select interesting frames and/or video excerpts for movie previewing on Video on Demand web sites.

Movie trailers are intended to raise a viewer's interest in a movie. As a consequence, they will not be a topical summary of the video, and they are likely to be constituted by 'teaser material' that should make a viewer curious to watch more.

In our approach to this problem, we originally were interested in assessing whether 'interestingness' could relate to salient narrative elements in a trailer. In particular, we wondered whether criteria for connecting production music fragments to storylines [5] would also be relevant factors in rater assessment of interestingness.

However, the rating acquisition procedure for the task did not involve full trailer watching by the raters, but rather the rating of isolated pairs of clips or frames. As such, while ideas in [5] largely considered the dynamic unfolding of a story, a sense of overall storyline and longer temporal dynamics could not be assumed in the current task.

We ultimately decided on pursuing a simpler strategy: the currently presented approaches investigate to what extent *the clear presence of people*, as approximated by automated face detection results, indicate visual environments which are more interesting to a human rater. The underlying assumption is that close-ups should attract a viewer's attention, and as such may cause larger empathy with the shown subject or its environment. It will be interesting to consider to what extent this currently proposed heuristic method will compare against more agnostic direct machine learning techniques on the provided labels.

MediaEval 2016 Workshop, Oct. 20-21, 2016, Hilversum, Netherlands.

Data	MAP
video ground truth on image set	0.1747
image ground truth on video set	0.1457

Table 1: MAP values obtained on development set by swapping ground truth annotations of image and video.

2. CONSIDERATIONS

In designing our current method, several considerations coming forth from the task setup and provided data were taken into account.

First of all, interestingness assessments only considered pairs of items originating from the same trailer. Therefore, given our current data, scored preference between items can only meaningfully be assessed *within the context of a certain trailer*. As a consequence, we choose to only focus on ranking mechanisms restricted to a given input trailer, rather than ranking mechanisms that meaningfully can rank input from multiple trailers.

Secondly, the use case behind the currently offered task considered helping professionals to illustrate a Video on Demand (VOD) web site by selecting interesting frames and/or video excerpts of movies. The frames and excerpts should be suitable in terms of helping a user to make a decision on whether to watch a movie or not. As a consequence, we assume that selected frames or excerpts should not only be interesting, but also representative with respect to the movie's content.

Thirdly, the trailer is expected to contain groups of shots (which may or may not be sequentially presented) originating from the same scenes.

Finally, binary relevance labels were no integral part of the rating procedure, but added afterwards. As a consequence, finding an appropriate ranking order will be more important in relation to the input data than providing a correct binary relevance prediction.

When manually inspecting the ground truth annotations, we were struck by the inconsistency between ground truth rankings on the Image Subtask vs. that obtained for the Video Subtask. To quantify this inconsistency, given that annotations were always provided considering video shots as individual units (so there were as many items considered per trailer in the Image Subtask as in the Video Subtask), we mimicked the evaluation procedure for the case ground truth would be swapped. In other words, we computed the MAP value for the Image Subtask in case the ground truth of the Video Subtask (including confidence values and binary relevance indications) would have been a system outcome, and vice versa. Results are shown in Table 1: it can be noted the MAP values are indeed not high. As we will discuss at the end of the paper, this phenomenon will be interesting to investigate further in future continuations of the task.

3. METHOD

As mentioned, we assess interestingness on the basis of (clearly) visible people. We do this for both Subtasks, and simplify the notion of 'visible people' by employing face detection techniques. While these techniques are not perfect (and false negatives, or missed faces, are prevalent), it can safely be assumed that when a face is detected, the face will be clearly recognizable to a human rater.

Both for the Image and Video Subtask, we follow a similar strategy, which can be described as follows:

1. Employ face detectors to identify those image frames that feature people. For each of these, store bounding boxes for all positive face detections.

2. In practice, the amount of frames with detected faces is relatively low. Assuming that *frames in which detected faces occur are part of scene(s) in the trailer which are important* (and therefore may contain representative content of interest), we consider the set of all frames with detected faces, and calculate the mean HSV histogram $\overline{H_f}$ over it.

3. For each shot s in the trailer, we consider its HSV histogram H_s and calculate the histogram intersection between H_s and $\overline{H_f}$ as similarity value:

$$sim(H_s, \overline{H_f}) = \sum_{i=0}^{|\overline{H_f}|-1} \min(H_s(i), \overline{H_f}(i)).$$

4. Normalize the similarity scoring range to the $[0, 1]$ interval to obtain confidence scores. The ranking of shots according to these scores will be denoted as `hist`.

5. Next to considering histogram intersection scores, for each shot, we consider the bounding box area of detected faces. If multiple faces are detected within a shot, we simply sum areas.

6. The range of calculated face areas also is scaled to the $[0, 1]$ interval.

7. For each shot, we take the average of the normalized histogram-based confidence score and the normalized face area score. These averages are again scaled to the $[0, 1]$ interval, establishing an alternative confidence score which is boosted by larger detected face areas. The ranking of shots according to these scores will be denoted as `histface`.

Both for the Image and Video Subtask, we submitted a `hist` and `histface` run. Below, we give further details on what feature detectors and implementation details were used per subtask.

3.1 Image Subtask

For the Image Subtask, each shot is represented by a single frame. The HSV color histograms for each frame are taken out of the precomputed features for the image dataset [4].

No face detector data was available as part of the provided dataset. Therefore, we computed detector outcomes ourselves, using the head detector as proposed by [7], and employing a detection model as refined in [6]. The features were computed employing the code released by the authors[1]. This head detector does not require frontal faces, but also is designed to detect profile faces and the back of heads, making it both flexible and robust.

[1] `http://www.robots.ox.ac.uk/~vgg/software/headmview/`

Run name	MAP
`image_hist`	0.1867
`image_histface`	0.1831
`video_hist`	0.1370
`video_histface`	0.1332

Table 2: MAP values obtained on development set.

Run name	MAP
`image_hist`	0.2202
`image_histface`	0.2336
`video_hist`	0.1557
`video_histface`	0.1558

Table 3: Official task evaluation results: MAP values obtained on test set.

We sort the obtained confidence values, and apply an (empirical) threshold to set binary relevance. For the `hist` run, all items with a confidence value higher than 0.75 are deemed interesting; for the `histface` run, the threshold is set at 0.6.

3.2 Video Subtask

For the Video Subtask, in parallel to our approach for the Image Subtask, we consider HSV color histograms and face detections. For this, we can make use of released precomputed features. However, in contrast to the Image Subtask, these features now are based on multiple frames per shot.

In case of the HSV color histograms [4], we take the average histogram per shot as representation. For face detection, we use the face tracking results based on [1] and [2], and consider the sum of all detected face bounding box areas per shot.

The binary relevance threshold is set at 0.75 for the `hist` run, and at 0.55 for the `histface` run.

4. RESULTS AND DISCUSSION

Results of our runs as obtained on the development and test set are presented in Tables 2 and 3, respectively. The results on the test set constitute the offical evaluation results of the task.

Generally, it can be noted that MAP scores are considerably lower for the Video Subtask than for the Image Subtask. Also looking back to the results in Table 1, it may be hypothesized that the Video Subtask generally is more difficult than the Image Subtask. We would expect for temporal dynamics and non-visual modalities to play a larger role in the Video Subtask; aspects we are not considering yet in our current approach.

When comparing the obtained MAP against the scores seen in Table 1, we notice that our scores are comparable, or even better. Furthermore, comparing results for the test set vs. the development set, we see that scores slightly improve for the test set, suggesting that our modeling criteria were indeed of certain relevance to ratings in the test set.

For future work, it will be worthwhile to further investigate how universal the concept of 'interestingness' is, both across trailers, and when comparing the Image Subtask to the Video Subtask. The surprisingly low MAP scores when exchanging ground truth between Subtasks may indicate that human rater stability is not optimal, and/or that the two Subtasks are fundamentally different from one another. Furthermore, as part of the quest for a more specific definition of 'interestingness', a continued discussion on how interestingness can be leveraged for a previewing-oriented use case will also be useful.

5. REFERENCES

[1] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 2005.

[2] M. Danelljan, G. Häger, F. Shahbaz Khan, and M. Felsberg. Accurate scale estimation for robust visual tracking. In *Proceedings of the British Machine Vision Conference*. BMVA Press, 2014.

[3] C.-H. Demarty, M. Sjöberg, B. Ionescu, T.-T. Do, H. Wang, N. Q. Duong, and F. Lefebvre. MediaEval 2016 Predicting Media Interestingness Task. In *Proc. of the MediaEval 2016 Workshop*, Hilversum, The Netherlands, October 2016.

[4] Y.-G. Jiang, Q. Dai, T. Mei, Y. Rui, and S.-F. Chang. Super Fast Event Recognition in Internet Videos. *IEEE Transactions on Multimedia*, 177:1–13, 2015.

[5] C. C. S. Liem, M. A. Larson, and A. Hanjalic. When Music Makes a Scene — Characterizing Music in Multimedia Contexts via User Scene Descriptions. *International Journal of Multimedia Information Retrieval*, 2:15–30, 2013.

[6] M. Marin-Jimenez, A. Zisserman, M. Eichner, and V. Ferrari. Detecting People Looking at Each Other in Videos. *International Journal of Computer Vision*, 106(3):282–296, February 2014.

[7] M. Marin-Jimenez, A. Zisserman, and V. Ferrari. Here's looking at you, kid." Detecting people looking at each other in videos. In *British Machine Vision Conference*, 2011.