Personalized Egocentric Video Summarization for Cultural Experience

Patrizia Varini, Giuseppe Serra and Rita Cucchiara
Dipartimento di Ingegneria "Enzo Ferrari", Università degli Studi di Modena e Reggio Emilia
Via Vignolese, 905/b, Modena MO 41125, Italy
name.surname@unimore.it

ABSTRACT

Recent egocentric video summarization approaches have dealt with motion analysis and social interaction without considering that user can be interested in preserving only part of the video related to his interests. In this paper we propose a new method for personalized video summarization of cultural experiences with the goal of extracting from the streams only the scenes corresponding to a user's specific topics request, chosen among the shots in which it's possible to deduce that the visitor was focusing on a point of interest. Preliminary experiments show that our approach is promising and allows visitor to better customize the summary of his experience.

Categories and Subject Descriptors

I.4 [**Image processing and computer vision**]: Applications

General Terms

Algorithms, Design, Experimentation

Keywords

Video summarization, egocentric vision, wearable devices

1. INTRODUCTION

Video summarization is gaining an increasing attention in multimedia and vision research community as humans need a fast growing support to analyze, crawl and search specific contents in the large and continuously swelling amount of digital video streams available on the internet.

In current literature, most methods for video summarization are based on visual features analysis, visual saliency or video quality assessment. Potapov *et al.* [6] achieve category specific video summarization, first obtaining temporal segmentation into semantically-consistent segments, delimited not only by shot boundaries but also general change points, then apply a SVM classifier to each segment to assigns category relevance score to each segment. Jain *et al.* [1] recognize shared content in videos and joint attention by finding the number of overlapping 3D static points.

ICMR '15, June 23 - 26, 2015, Shanghai, China

http://dx.doi.org/10.1145/2671188.2749343.

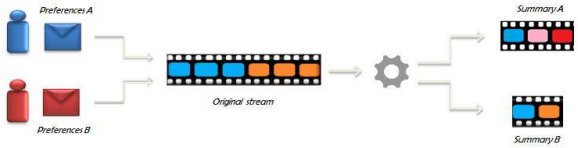

Figure 1: User requests personalized summarization

Recently the wide spreading use of head-mounted cameras has made popular life-logging video. Typically this egocentric videos consist of very long streams of data with a ceaseless jumping appearance, very frequent changes of observer's focus and lack of hard cuts between scenes, thus requiring new methodologies.

Lee *et al.* [4] proposed a egocentric video summarization method that focuses on learning importance cues for each frame, such as objects and people the camera wearer interacts with. In particular, they measure importance on a combination of interaction distance, gaze, object-like appearance and motion and likelihood of a person's face within a region.

Lu and Grauman [5] handle egocentric video summarization partitioning videos into sub-shots on the basis of motion features analysis, smooth the classification with a MRF and then select a chain of sub-shots choosing the ones in which they can detect the reciprocal influence propagation between important objects and characters. Yeung *et al.* [7] present techniques to evaluate video summarization through text, by measuring how well a video summary is able to retain the semantic information contained in its original stream making use of textual summarization benchmarking tools.

Although these summarization techniques deal with egocentric characteristics, they do not take into account the particular user's preferences. In fact, a user can prefer to remember some events rather than others (see Fig. 1). For example, some users might be concerned only with romanesque and baroque art, and might not be interested in preserve shots related with modern art or other kinds of experiences (food, shopping etc.), viceversa users which might be interested only in contemporary architecture might like to discard shots related to ancient art.

In this paper we present a method of user personalized egocentric video summarization in a cultural experience sce-

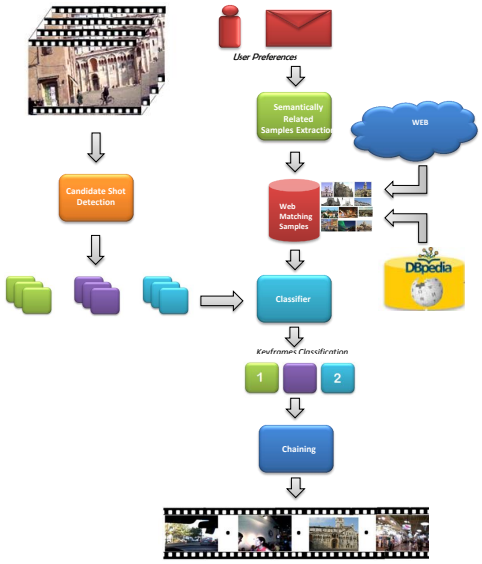

Figure 2: Schematization of the proposed method

nario. The approach relies on extracting from the stream only the scenes, chosen among specific shots assembled with groups of frames that can be put in relation with an observer's attention behaviour, matching some questioner user's declared topics of interest, so that different users typically obtain different summaries from the same stream. User's preferences are represented by different groups of keywords. Since user's declared topics of interest can not be constrained, we propose a data-driven approach that extracts positive and negative training samples from the web. Our preliminary experimental results show that this approach is able to exploit user's preference to obtain a personalized summarization of a cultural visit video.

2. PERSONALIZED SUMMARIZATION

Our approach can be roughly described as the set of two main tasks (see Fig. 2). The first aims to identify the candidates of the relevant scenes, discarding all the groups of frames related to irrelevant experiences or in which the observer is changing his focus of attention, and ends with the candidate shots detection and keyframes extraction. Discarding the non relevant shots and limiting the analysis to keyframes, aims to reduce the computational overhead and to focus on the research of presumably relevant features. Our hypothesis relies on the assumption, tailored on a typical cultural experience scenario, that the relevant scenes are associated to a camera's viewer behavior due to the presence of attention patterns. The second task aims to extract from the candidate shots only the ones that maximize the score of semantic relatedness to the preferences requested by the user and the visual diversity. To achieve this goal we build specific classifiers of the topics of interest and find out the scenes that achieve the highest score. In order to identify reliable image training samples from the web, we evaluate importance on semantic relatedness with user input using DBpedia and visual difference.

2.1 Candidate shot detection

We are interested in identifying and extracting from the original stream shots that can be put in relation with the due of camera wearer "paying attention" pattern. In general, in a cultural experience scenario visitor can have different behaviours, but we focus on four attitude patterns: transit from one point of interest to another, changing the focus of attention, paying attention to something which is relevant, or wandering around without showing signs of interest. Of course behaviours are not directly detectable but we put them in relation with detectable motion patterns from video analysis.

Thus, we define the following observable motion classes: "Static", "Walking", "Higher speed" (running, jumping, falling, etc), "On wheels" (driving, on bus, etc.), "Head Rolling" and "Head Pitching".

Therefore we model camera's wearer pattern behaviours as a Hidden Markov Models with the four behaviour states and the six directly observables defined upper.

In a general Hidden Markov Model with N hidden states and M observables, the model is completely described by the initial state probability, the transition $N \times N$ matrix with elements $a_{ij} = p(q_{t+1} = j | q_t = i), 1 \leq i, j \leq N$ and emission probabilities $N \times M$ matrix with elements $b_{jk} = p(o_t = \theta_k | q_t = j), 1 \leq j \leq N, 1 \leq k \leq M$.

The transition matrix pdf was initialized assumimg that the "transit" and "wandering around" states were the most frequent, the emission matrix was initialized evaluating the probability to observe the emission o_t in the q_t state.

Once defined the model, we have used Baum-Welch algorithm for training to obtain a better detection of the interest sequences of events. The models were fed with observables vector and probability of precedent state and Viterbi algorithm was iterated to specialize the parameters.

To classify the observable primitive classes, we analyze frame quality assessment features and motion pattern by partitioning frame using a 3×3 grid. In particular blurriness is used to assess quality frame. We compute this feature by using the method proposed by Roffet *et al.* [2]. They assume that the sharpness of an image is contained in its gray component and estimate the blur annoyance only on the luminance component, computing and evaluating the line and row difference between the original image and the image obtained applying to it a horizontal and a vertical strong low-pass filter. The blurriness descriptor is obtained by concatenating sector features.

Motion feature is based on optic flow and its gradient histograms estimated using the Farneback algorithm. Considering the optic flow computed for each couple of consecutive frames, the relative apparent velocity and acceleration gradient of each pixel is V_x, V_y, A_x and A_y. These values are expressed in polar coordinates as in the following:

$$M_V = \sqrt{V_x^2 + V_y^2} \qquad \theta_V = \arctan(V_y / V_x) \tag{1}$$

$$M_A = \sqrt{A_x^2 + A_y^2} \qquad \theta_A = \arctan(A_y / A_x) \tag{2}$$

We compute a histogram by concatenating the magnitudes M_V and M_A (quantized in eight bins), with the orientations θ_V and θ_A, (quantized in eight bins) weighting them by their magnitude respectively.

In order to reduce the jumpy values of motion measures due to meaningless head motion, the feature vector descriptors have been averaged over a window of about 20 frames (when acquiring at 29 FPs). This window corresponds to a

duration of less than a second and it has been considered to be a reasonable compromise to reduce randomness without information loss. In fact, the typical interval duration of head movement in the "paying attention" pattern has been put in relation to visual fixation, studied using gaze analysis, that is about 330 ms [3] but has a wide range of variation. To speed up classification task, a linear multiclass SVM has been trained over the six identified classes.

2.2 Semantic classification and Shot chain

To identify the set of candidate shots that maximize the relatedness to the user's preferences we build a visual recognition system based on discrete classifiers. Since topics of interest requested by the user can be infinite, visual classification based on a number of rigidly defined classes is not feasible. To deal with this problem we proposed a data-driven approach that gathers positive and negative training samples from the web. To obtain a reliable training set, we analyze importance on semantic relatedness with user's preference and we expand semantically them using DBpedia.

In fact, DBpedia semantic network has become a new enabling resource for semantic processing and natural language processing, with its wide spread semantic coverage, its constant updates and its semantic support structure, taxonomies (hyponyms, hyperonyms, synonyms, antonyms), translations for each lexical word, cross references between related topics, disambiguation pages, ontology management and topic inference.

We regard Dbpedia as a undirected weighted graph $G = \{V, E\}$ where $V = \{1, ...n\}$ are the nodes representig concepts and $E \subset V \times V$ are the edges representing the links among nodes. To detect semantic community in DBpedia we use the recursive Girvan-Newman algorithm. The algorithm starts with computing the "betweenness" score for each of the edges ("betweenness" of an edge is the number of shortest paths between pairs of nodes that run along it). Remove the edge with the highest score. Compute the betweenness of all edges affected by the removal. The last two steps are repeated until no edges remain.

At last for each detected semantic community including the user keywords and at most other K terms (in our experiments we fixed K=3), we evaluate average of the shortest paths between communities members and the geolocalized place where the video has been captured using Dijkstra algorithm. The basic intuition is that semantic concepts that are strictly related with user's preferences and visit location can improve the search terms for collecting training images.

Let's suppose a user's preference is "Reinessance" and the video is captured in Ferrara city. Fig 3 shows the positive examples, selecting images with high relevance score from a google image search engine, and negative samples randomly taken. With our approach (see Fig. 4) this user's preference is expanded in the semantic community "Reinessance Ferrara Diamond-palace". Therefore positive samples are extracted from a image search on this set of terms, explicitly excluding all images labeled with semantic concepts or tags that have a shortest path distance from the expanded preference over a threshold. Negative samples are gathered using a search of semantic concepts reached moving on the graph from the expanded preference of N steps (we empirically fix it to ten). Finally, starting from the positive and negative samples extracted we build semantic classifiers using the Bag of Words approach (BOW).

Figure 3: Example of positive and negative samples without geolocalization and semantic expansion (keyword: Reinessance. Location: Ferrara)

Figure 4: Example of positive and negative samples with the proposed geolocalization and semantic expansion. (keyword: Reinessance. Location: Ferrara)

Relevance of each shots is computed taking into account classification scores (S) and visual diversity (D):

$$R(s) = w_1 S + w_2 D \tag{3}$$

For each shot, S is computed as the sum of the scores obtained on each keyframe by all classifiers learned form the expanded preference communities and normalized by shot length. To measure visual diversity D, we represent a shot as a phrase (string) formed by the concatenation of the bag-of-words representations of consecutive characters (keyframes). To compare these phrases (or shots) we use the Needleman-Wunsch distance defined as the number of operations required to transform one string into the other. In particular, D is the normalized sum of the distances of the shot with respect to the adjacent ones. Based on preliminary experiments, we empirically fix the weighting coefficients w_1 and w_2.

3. EXPERIMENTAL RESULTS

To evaluate the performance of our approach we collected ten videos captured by tourists that spend some time to visit cultural cities. Each video is about one hour long and taken in a uncontrolled setting. They show the experience visitors such as a visit of cultural interest point (church, monument etc), shopping or walking. The camera is placed on the tourist's head and captures a 720 × 576, 29 frames per second RGB image sequence.

A subset of 7200 annotated frames is used in order to test our methodology to recognize the motion classes: "Static", "Walking", "High speed", "On wheels", "Head Roll" and "Head Pitch". First, we examine the effectiveness of our feature vector representing frame quality assessment features and motion pattern.

In Table 1 we compare average class accuracy of our results to the features proposed by Lu *et al.* [5] (based on blurriness, optical flow orientations weighted over magnitudes and magnitudes). The Figure 5 shows the performance of the two feature descriptors per class.

As can be see our descriptor achieves a better performance. In fact, adding descriptors related to apparent ac-

	Lu *et al.* [5]	**Our approach**
Accuracy	62.92	**72.48**

Table 1: Comparision of classification accuracy.

Figure 5: Classification accuracy using different descriptors: a) feature vector proposed by Lu *et al.* [5]; b) our feature vector.

celeration can help to distinguish abrupt and random movements of the head from other kinds of motion within the frame since these features are more sensible with respect to the apparent velocity.

In addition, we evaluate the ability of our approach to collect reliable training set analyzing the semantic relatedness of user's preference with DBpedia knowledge structure. We show the results of two experiments where the user's preferences are expressed by keywords "Romanesque" and "Baroque" for the first one and by "Picture cards" and "Clothes Shopping"' for second. To analyze our results we compare the performance obtained by classifiers trained with web images extracted using the proposed geolocalization and semantic expansion with classifiers learned using only the user keywords. In particular, for each image we extract SIFT descriptors computed at four scales (4, 6, 8, and 10), over a dense regular grid with a spacing of 6 pixels. The codebook size is set to 2000. Images are hierarchically partitioned into 1×1, 2×2 and 4×4 blocks on 3 levels respectively. SVM classifiers have been trained on the collected images (60% for training and 40% validation and testing) and performance was evaluated using 10-fold crossvalidation. Fig. 6 shows the classification comparison between our approach and the baseline in term the F1-score. Notice that in all cases classification performances outperform the baseline. In particular, we observe that the information about visit location can better restrict the visual appearances of the topics of interest requested by the user.

Finally, we perform a "blind taste test" in which, for each video of the dataset, the summarization based on our approach and a baseline are shown to eight students, that have to report which summary best meets the user's preferences related to video. We first show to the students a browsable sped-up version of the entire original videos, and ask them to write down the shots in which they think the users's preferences given as sample input are fit. Afterward, for each original video, we show them two summaries: one is obtained with the proposed method, the other is from a baseline method in which all candidate shots are chained. We do not reveal the order as it is randomly obtained. After viewing both, the subject is asked, "Which summary better shows the relevant events taking into account the user's

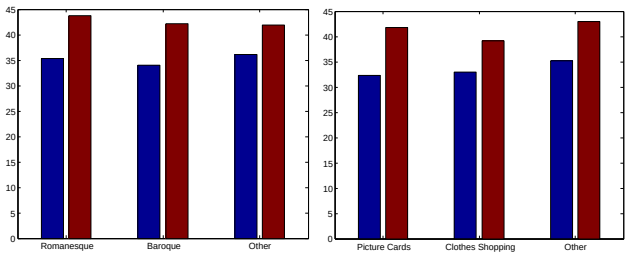

Figure 6: Classification comparison between classifiers trained with only the user's preference (blue hist.) and with proposed geolocalization and semantic expansion (red hist.).

preferences?". We used a Likert scale with a score between 1 and 5, where 1 was "no good summarization" and 5 "perfect summarization". This test shows that 84% of the comparisons assigns a higher score to summaries obtained with our approach with respect to the baseline.

4. CONCLUSIONS

In this paper we have introduced a novel approach to user personalized egocentric video summarization in a cultural experience scenario. The approach focused on extracting from the original video the relevant shots with a "paying attention" pattern. These candidate shots are further filtered in order to obtain a summary matching the requested user preferences. Our preliminary results show that the proposed approach is able to take into account user's preference in order to obtain a personalized summarization.

Acknowledgments

This work was partially supported by the Fondazione Cassa di Risparmio di Modena project "Vision for Augmented Experiences" and the PON R&C project DICET-INMOTO (Cod. PON04a2_D).

5. REFERENCES

[1] S. Antani, R. Kasturi, and R. Jain. A survey on the use of pattern recognition methods for abstraction, indexing and retrieval of images and video. *Pattern Recognition*, 35(4):945 – 965, 2002.

[2] F. Crété-Roffet, T. Dolmiere, P. Ladret, M. Nicolas, et al. The blur effect: Perception and estimation with a new no-reference perceptual blur metric. In *Proc. of SPIE*, 2007.

[3] J. M. Henderson. Regarding scenes. *Current Directions in Psychological Science*, 16(4):219–222, 2007.

[4] Y. J. Lee, J. Ghosh, and K. Grauman. Discovering important people and objects for egocentric video summarization. In *Proc. of CVPR*, 2012.

[5] Z. Lu and K. Grauman. Story-driven summarization for egocentric video. In *Proc. of CVPR*, 2013.

[6] D. Potapov, M. Douze, Z. Harchaoui, and C. Schmid. Category-specific video summarization. In *Proc. of ECCV*, 2014.

[7] S. Yeung, A. Fathi, and L. Fei-Fei. Videoset: Video summary evaluation through text. *CoRR*, abs/1406.5824, 2014.