Egocentric Video Summarization of Cultural Tour based on User Preferences

Patrizia Varini Giuseppe Serra Rita Cucchiara

Dept. of Information Engineering - University of Modena and Reggio Emilia
Via Pietro Vivarelli, 10 (Int. 1) 41100 Modena, Italy
name.surname@unimore.it

ABSTRACT

In this paper, we propose a new method to obtain customized video summarization according to specific user preferences. Our approach is tailored on Cultural Heritage scenario and is designed on identifying candidate shots, selecting from the original streams only the scenes with behavior patterns related to the presence of relevant experiences, and further filtering them in order to obtain a summary matching the requested user preferences. Our preliminary results show that the proposed approach is able to leverage user's preferences in order to obtain a customized summary, so that different users may extract from the same stream different summaries.

Categories and Subject Descriptors

I.4 [**Image processing and computer vision**]: Costumized egocentric video summarization; H.3.1 [**Information systems**]: Content analysis and indexing

General Terms

Algorithms, Design, Experimentation

Keywords

Video summarization, egocentric vision, wearable devices

1. INTRODUCTION

The increasingly popular life-logging streams, captured by head-mounted cameras, are claiming new techniques for automatic analysis, understanding and summarizing, being characterized by continuous changes of observer's focus, incessantly changing objects appearance and lack of hard cuts between scenes.

Lee *et al.* [4] propose egocentric video summarization method that focuses on learning importance cues for each frame, such as objects and people the camera wearer interacts with, using features related with interaction distance, gaze, object-like appearance and motion and likelihood of a person's face within a region. Lu and Grauman [6] 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, recovering their story. 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. Although these summarization techniques deal with egocentric characteristics, they produce a univocal summary, not taking into account that different viewers, according to their own preferences, might prefere to retain some events rather than others. For instance, art lovers and fashion enthusiasts might want to extract different summaries from the same stream. In this paper we tackle customized ego-video summarization in a cultural heritage scenario, extracting from the streams only the scenes matching user's specific topic requests. In particular, to identify candidate relevant shots, we propose a behavior pattern detection classifier based on motion related features, as visual velocity, visual acceleration gradient and 3D GPS detected velocity. Due to user preferences heterogeneity, semantic classifiers for requested topics are built using a data-driven approach, that exploits geolocalization information and DBPedia knowledge. Experimental results show that our approach is able to achieve a customized summarization w.r.t. the user topic requests.

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

DOI: http://dx.doi.org/10.1145/2733373.2806367.

2. CUSTOMIZED SUMMARY

Our method takes a long first person video and user preferences as input and returns a customized short video summary as output. First, we identify the candidate subshots of video, discarding all the groups of frames whose motion based pattern cannot be put in relation with the presence of a relevant experience for the observer, for instance when he's changing his focus of attention. In particular our hypothesis relies on the assumption, tailored on a typical cultural experience scenario, that relevant scenes are associated to a camera's viewer behavior due to the presence of attention patterns. Discarding the non relevant subshots and limiting the analysis to keyframes, aims to reduce the computational overhead and to focus on the analysis of presumably relevant experiences. We assume that behavior patterns can be put in relation with user motion patterns, described by

directably measurable features like frame visual assessment and apparent 2D and real 3D GPS measured motion. Thus we define a motion taxonomy, structured in two tiers, "Body in motion" and "Body still". In "Body still" tier we find the classes "Static" (Body and head stand still) and "Looking around" (Body is still, head is moving). In the "Body in motion" tier we find "Walking" (Body is walking, head is approximately still), "Running" (Body running and Head in coherent motion), "On wheels" (Body and Head are still respect to a moving wheel mean of transport), "Wandering" (Body is in motion and head is rolling and/or pitching). To detect these classes, we analyze the aforectited features by partitioning frame in a 3×3 grid. Finally we smooth the motion classification with a Hidden Markov Model to shape, from the primitive motion classes, the behavior patterns defined as "Attention" (where user is paying attention to something), and the homonym w.r.t. motion classes "Looking around","Walking", "Running", "On wheels", "Wandering", filter the "Attention" behavior shots, that represent candidate shots, and extract the keyframes. Father, we select from the candidate subshots only the ones that maximize the score of semantic relatedness to the user preferences and the visual diversity. Assuming that in Cultural Heritage scenario we usually deal with well localized classes of objects, we expect to enhance our performances leveraging GPS coordinates to gain context awareness. Thus, we build specific classifiers for the topics of interest using a data driven approach, to extract "on the fly" reliable image training, location relevant, samples from the web, evaluating importance on concept relatedness with user input using DBpedia semantic knowledge. Figure 1 sinthesizes our method.

2.0.1 Visual Motion Descriptor

Visual motion feature is based on optic flow and acceleration gradient histograms estimated using the Farneback algorithm on frame sections using a 3×3 grid. 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:

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

Acceleration gradients on x and y are computed for horizontal and vertical component and :

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

We compute the motion histogram by concatenating the apparent motion magnitudes M_V and M_A, with the orientations θ_V and θ_A, both quantized in eight bins for each frame section, weighting them by their magnitude respectively.

To assess the frame quality, we compute blur feature by using the method proposed by Roffet *et al.* [1]. The blurriness descriptor is obtained by concatenating sector features.

2.0.2 3D Motion Feature

Visitor's velocity and stops are a semantically relevant part of a touristic visit, being related to his intention and interests. Thus, collecting spatial coordinates via GPS tracking every second, allows to compute real velocity in three dimensions. Therefore visitor's trajectories are represented by movement tracks, consisting in the temporal sequence of the spatio-temporal points, meant as pairs compound with coordinate in space and in time $\{p_i = (x_i, y_i, z_i, t_i)\}$, where $(x_i, y_i, z_i) \in \mathcal{R}^3, t_i \in \mathcal{R}^+$ for $i = 0, 1, ..., N$ and $t_0 < t_i < t_N$. Even if trajectory is known, it does not entirely embed insight about stops and moves semantic information, infact different visitors may have the same velocities in a number of subsequent points, but due to their different motion patterns, these points may have different classification outcomes for each of them. So we propose to adopt a spatio-temporal clustering algorithm to add the stop or move further information to our overall motion descriptor. K-means is a standard and efficient clustering algorithm, but needs to calculate the number of clusters, instead, we propose the use of a Shared Nearest Neighbor (SNN) density-based algorithm [2], whose extension in 4 spatio-temporal dimensions was first explored by [5], that is able to deal with noisy clusters of different densities, sizes and shapes. SNN relies on strength or similarity concept, evaluated on the number of nearest neighbors that couples of points, belonging to a set of N points in a metric space D, share. Therefore, our final overall 3D motion descriptor is obtained chaining the 3D real velocity components measured every second with a boolean value that takes into account if the corresponding point must be regarded as a stop or a move.

2.0.3 Subshot Scoring

In order to reduce the jumpy values of motion measures due to meaningless head motion, the aforecited feature vector descriptors have been averaged over a window of 25 frames. This window corresponds to a duration of less than a second (acquiring at 29 FPs) and has been regarded 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. Finally, we compose our overall feature vector concatenating the apparent motion histogram described in 2.0.1, averaged over the 1 sec window as described upperline, with the three components of real velocity histogram and a boolean value that indicates if the frame, within the window, belongs to a pattern of stop or of move of the visitor.

To speed up classification task, a linear multiclass SVM has been trained over the six identified motion classes and, using a Hidden Markov Model, the six behavior states are obtained from the six motion observables defined upper. Candidate relevant shots are finally identified as the ones associated with "Attention behavior" pattern.

2.1 Semantic classification and Shot chain

To identify the set of candidate shots that maximize the relatedness to the user's preferences extracted on input keywords, we build a visual recognition system based on discrete classifiers. Since topics of interest requested by the user can be potentially limitless, visual classification based on a number of rigidly defined classes may lead to poor results. To deal with this problem we proposed a data-driven approach that gathers 100 positive and negative training samples from the web, analyzing visual importance on semantic relatedness with user's preferences.

DBpedia is an important dataset in Linked Open Data, being provided with constantly updated semantic support structure, taxonomies (hyponyms, hyperonyms, synonyms,

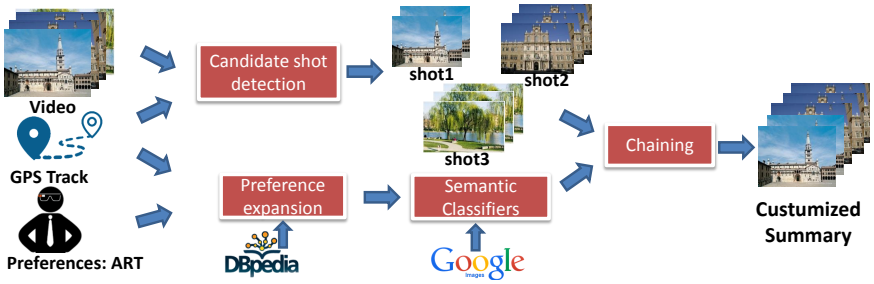

Figure 1: Schematization of the proposed method

antonyms), 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 representing 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). Then edges with the highest score are removed and the betweenness of all edges affected by the removal are computed. The last two steps are repeated until no edges remain.

At last for each detected semantic community including the user keywords and at most three related concepts, whose edge weight is over a threshold, 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, being Cultural Heritage items highly location-specific. Therefore positive samples are extracted from an 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 preference cloud 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 (empirically fixed to ten). Finally, starting from the positive and negative samples extracted, we build semantic classifiers using the Bag of Words approach (BOW). Relevance of each shots is computed taking into account classification scores (S) and visual diversity (D): $R(s) = w_1 S + w_2 D$. 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 twelve videos captured by tourists that spend some time to visit cultural cities. Each video is about thirty minutes 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, 25 frames per second RGB image sequence. A subset of 26100 annotated frames is used in order to test our methodology to recognize the motion classes belonging to the taxonomy cited above.

3.0.1 *Shot classification according to visual and motion pattern*

We examine the effectiveness of our feature vector representing frame quality assessment features and motion pattern, both apparent, due to first person capture, and absolute, measured by GPS. First, we compare our Visual motion descriptor (VMD), with a recent similar feature vector proposed by Lu *et al.* [6] that exploits HOF and Blur. As can be seen in table 1, our VMD feature achieves a better performance. This may be mainly due to the presence of the acceleration gradient, that helps identify visual abrupt motion due to head movements. Moreover, we also present the performance of our final descriptor VMD-3DM (that includes 3D motion information).

In Table 1 we compare average class accuracy of the three approaches. Figure 2 shows the confusion matrix for VMD and VMD-3DM features.

Visual descriptor	Accuracy
Lu *et al.* [6]	61.7
VMD	69.7
VMD-3DM	74.1

Table 1: Comparision of classification accuracy.

As can be see our descriptor, with the addiction of GPS coordinates and real motion features, achieves a better performance. In fact, adding descriptors related to 3D motion can help to easily distinguish higher speed motions from wide head motion or sprawl movements. For this reason, among the classes that increase their accuracy, can be found found "Running" and above all "Moving on Wheels". The earning of "Looking around" class also may be due to the difficulty to discriminate steady head motion from sprawl behavior in motion in absence of measurements related to real motion.

3.0.2 *Keyframe classification according to relatedness to user preferences*

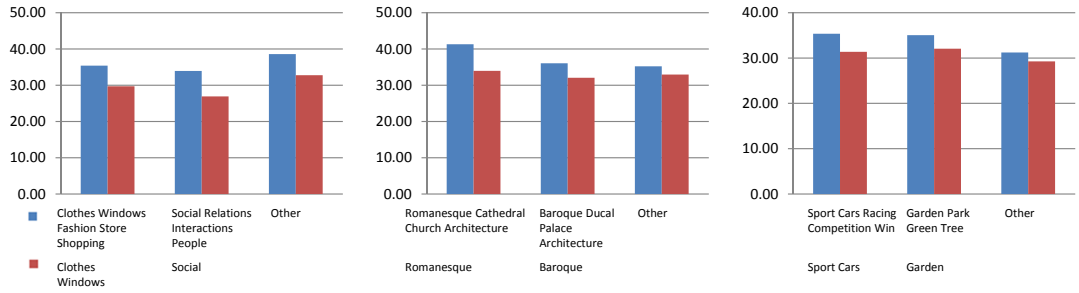

Figure 3: Keyframes classification accuracy with (blue) and without (red) semantic expansion

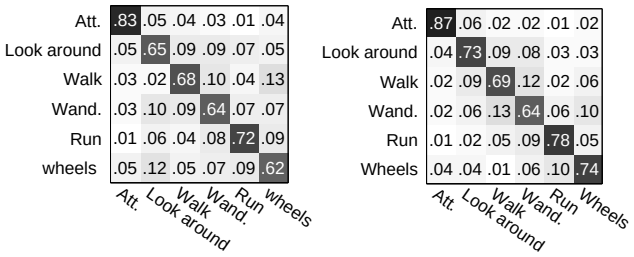

Figure 2: Classification accuracy using different descriptors: a) feature vector based on visual features (Blur, Optical Flow and Acc. Gradient; b) our feature vector.

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 three experiments, where the user's preferences are expressed by two groups of keywords: "Social" and "Clothes" for the first one, "Romanesque" and "Baroque" for the second and "Sport Cars" and "Gardens" for the third. 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. 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.

3.0.3 User experience

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 six 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 annotate the shots that they think are fitting the user's preferences. 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 a random selection of a fixed number of candidate shots are chained. We do not reveal the order as it is obtained randomly. After viewing both, each of them is asked to report which summary better matches the user's preferences in his opinion. 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 76% of the comparisons assigns a higher score to summaries obtained with our approach w.r.t. the baseline.

4. CONCLUSIONS

In this paper we have introduced a novel approach to obtain customized egocentric video summaries in Cultural Heritage scenario. The approach relied on detecting candidate shots, extracting from the original video the ones with a "paying attention" pattern, and further filtering them in order to obtain a summary matching the requested user preferences. Preliminary experiments show that our results are promising in enabling users to achieve personalized summary from visitor videos.

5. REFERENCES

[1] 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.

[2] L. Ertöz, M. Steinbach, and V. Kumar. Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data. In *Proc. of SDM*, 2003.

[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] Q. Liu, M. Deng, J. Bi, and W. Yang. A novel method for discovering spatio-temporal clusters of different sizes, shapes, and densities in the presence of noise. *International Journal of Digital Earth*, 7(2):138–157, 2014.

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

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