Egocentric Video Personalization in Cultural Experiences Scenarios

Patrizia Varini, Giuseppe Serra and Rita Cucchiara

Università degli studi di Modena e Reggio Emilia, Modena, Italy,
Via Pietro Vivarelli, 10
[name.surname]@unimore.it,
http://imagelab.ing.unimore.it

Abstract. In this paper we propose a novel approach for egocentric video personalization in a cultural experience scenario, based on shots automatic labelling according to different semantic dimensions, such as web leveraged knowledge of the surrounded cultural Points Of Interest, information about stops and moves, both relying on geolocalization, and camera's wearer behaviour. Moreover we present a video personalization web system based on shots multi-dimensional semantic classification, that is designed to aid the visitor to browse and to retrieve relevant information to obtain a customized video. Experimental results show that the proposed techniques for video analysis achieve good performances in unconstrained scenario and user evaluation tests confirm that our solution is useful and effective.

Keywords: Video analysis, Video personalization, Cultural Heritage

1 Introduction and related work

In recent years the widespreading use of wearable cameras to capture everyday life activities such as sport, education, social interactions and cultural heritage visits, has made popular egocentric videos. Typically they consist of long streams of data with a ceaseless jumping appearance, frequent changes of observer's focus and lack of hard cuts between scenes, thus requiring new methodologies for automatic analysis and understanding. There is a sharply increasing need of automated tools able to classify, search and select from these extremely long and continuous life logging streams, only the most relevant scenes according to the user preferences and to the specific purpose, eventually enriching them with customized semantically related content.

Various approaches exist for data visualization to help users navigation in the selected videos. Visualization systems based on timeline slider and on shots sequence show are the most common and easiest way to get a quick overview of video content but they suffer by lack of semantic categorization and poor scalability for large documents. Campanella *et al.* [2] propose a data visualization system to explore and annotate video sequences where contents are analyzed and displayed organized in classes and browsable in a feature distributed space

shown in a 2D Cartesian plane, where each axis corresponds to one feature type selected by the user and each shot is represented by a little square filled by the dominant colour of the shot. Snoek *et al.* [10] present the MediaMill video search engine, and proposes, among others, Sphere Browser, that represents a novel interface for searching through semantic space using conceptual similarity. This is obtained classifying shots with a similar conceptual index clustered together into threads. The Sphere Browser shows the timeline of the current video on the horizontal axis, and for each shot from the video it displays the relevant threads on the vertical axis. It uses a linear ordering to ranking video data. The vertical axis is related to a selected concept. The horizontal one is used to visualize video program in a timeline from which a keyframes is selected.

Moving to video personalization, Wei *et al.* [11] propose a novel architecture for video personalization and caching for resource constrained environments such as mobile devices, that performs automatic video segmentation and video indexing based on semantic video content, and generates personalized videos based on client preference using a Multiple-Choice Multi-Dimensional Knapsack Problem (MMKP)-based video personalization strategy. Araujo *et al.* [1] present a system for personalization of interactive digital media in educational environment, which combines context of access, user preferences and device presentation constraints in order to provide an interactive access experience. It allows content recommendation, ranking and personalization of interactive multimedia presentations captured in an instrumented classroom. These personalization techniques however do not take into account egocentric video peculiar issue. To best of our knowledge however, no one has addressed video personalization in egocentric vision.

Recently new methodologies related to egocentric video analysis have been developed to tackle its characteristic issues. Lee *et al.* [7] 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, using features related with gaze, object-like appearance and motion and likelihood of a person's face within a region. Lu and Grauman [9] 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.* [12] present a technique 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.

In this paper we propose a method for user egocentric video personalization which associates patterns of low level features to high level concepts relevant to different semantic levels, relying on geolocalization and on web dynamically extracted knowledge. We use a cultural experience scenario as use case, choosing candidate relevant semantic dimensions such as Points Of Interest (POI), visitor's behavior and spatial information about his stops and moves. Furthermore we present a web application that classifies and makes available shots corre-

sponding to different semantic levels, allowing the final user to select easily the relevant scenes, eventually according to his high level expressed preferences, expressed for sake of simplicity by simple groups of keywords containing names of classes (eventually with labels within the class to further filter the results, limited to the POI semantic level) belonging to one or more semantic levels. Our preliminary experimental results show that this approach is able to exploit dynamically user's preferences to obtain a personalized version of a cultural visit video.

2 Video personalization

We propose an approach for egocentric video personalization tailored on the use case of cultural experience scenario in which a video is segmented and classified in shots according to three different classes of semantic information: camera's wearer attitude or behaviour, stops and moves in the geolocalized traveled route and the presence of relevant cultural Points Of Interest.

In order to achieve a motion based classification of camera's wearer behaviour pattern, we define the underline motion taxonomy, structured in six classes. Annotations related to presence of stops and moves in the geolocalized trajectories are detected using a spatio-temporal clustering technique based on shared nearest neighbor. Detection of cultural POI is achieved by means of image classification using sets of positive and negative samples dynamically obtained from the web.

Observer's behavior pattern detection Based on the analysis of the visitor's typical behaviours, we define a taxonomy of a set of primitive motion classes: for the class "Person motion" the sub-classes "Static" (Body and head stand still), "Walking" (Body is walking, head is approximately still), "Higher speed motion" (Body running or jumping etc. and Head in coherent motion), "On wheels" (Body and Head are still respect to a moving on wheels mean of transport), for the class "Head motion" the sub-classes "Rolling" (Body is still or in motion and head is widely rolling) and "Pitching" (Body is still or in motion and head is widely pitching). To detect these classes, we analyze frame quality assessment and motion pattern features by partitioning frames 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.* [3], assuming 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 thus obtained by concatenating sector features.

Motion features are based on dense optic flow estimated using the Farneback's algorithm [5] and consist of optical flow and its gradient spatial histograms. Considering the optic flow computed for each couple of consecutive frames, the

relative apparent velocity and acceleration of each pixel is V_x, V_y, A_x and A_y. These values are expressed in polar coordinates as in the following:

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

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

For each of the 3×3 sections of the frame, 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 the magnitude respectively.

In order to smooth 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) as this has been regarded to be a reasonable compromise to reduce randomness without relevant information loss. In fact, the typical interval duration of head movement in the visual fixation pattern, studied using gaze analysis, is about 330 ms but has a wide range of variation [6]. Head movements themselves, measured with our approach, have been found to have a typical duration between 1 and 1.5 second (median 1,27 sec). To speed up classification task, a linear multiclass SVM has been trained over the six identified classes.

Based on the classification of these primitive classes, we exploit Hidden Markov Model to recognize the following behavior patterns and gain classification smoothing: attention, changing point of attention, wandering around and traveling from one point to another. In particular, we estimate the transition and emission probabilities from sample sequences in a supervised approach, in order to obtain a smoothed classification frame vector. A n-states Hidden Markov Model may be completely described by the initial state probability, by transition matrix from state S_i and by *pdf* matrix of observable O_i. Once defined the model the likelihood of the hidden state variables is computed with the Baum-Welsh algorithm which uses a forward and backward recursion. A model for each pattern of behaviour is prepared where states are related to different motion states. Afterward, the models are fed with observables vector and probability of precedent state and Viterbi algorithm is used to calculate the *pdf* to be higher than current, assigning the correspondent class.

Stop and move detection In cultural experience scenarios, stops are a semantically relevant part of a touristic visit, identified as places where a visitor has stayed for a minimum amount of time. Collecting the geographic locations by means of GPS sensors, trajectories are represented by movement tracks, that basically consist in the temporal sequence of the spatio-temporal points, meant as pairs compound with coordinate in space and in time $\{p_0 = (x_0, y_0, z_0, t_0), ..., p_N = (x_N, y_N, z_n, t_N)\}$, where $(x_i, y_i, z_i) \in \mathcal{R}^3, t_i \in \mathcal{R}^+$ for $i = 0, 1, ..., N$ and $t_0 < t_i < t_N$. As this definition itself does not embed any insight about stops and moves semantic informations, we proposed to adopt a spatio temporal clustering algorithm.

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 [4], whose extension in 4 spatio-temporal dimensions was first explored by [8], that is able to deal with clusters of different densities, sizes and shapes and with noise. 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, computed on the basis of a metric distance: $S(p,q) = size(kNN(p) \bigcap kNN(q))$. Density of a point p is evaluated as the number of points, within a radius Eps, defined so that its Eps-neighborhood is $N_{Eps}(p) = \{q \in D | dist(p,q) \leq Eps\}$, assumed that dist() is the Euclidean distance function. Then we define a cluster C as a set of elements in which for every point p there is at least a point q in C so that p is inside of the Eps-neighborhood of q and $N_{Eps}(q)$ contains at least a minimum number ($MinPts$) of points (q points are defined as core or representative points).

Thus assuming that a stop is semantically identified as the permanence of the visitor in a location (within a given radius) for a certain period of time, the used algorithm relies on fixing the number of nearest neighbors k, a density threshold $MinPts$ for a core point and a fixed radius (Eps), and starts with creating the similarity graph, reducing it to keep only the most similar nodes with their strength over the $MinPts$ threshold, discarding noise points as non-core points that are not within a radius Eps of a core point, and putting together in a same cluster core points within a Eps radius. Non-core points and non-noise points are classified as reachable points and assigned to clusters of their nearest core point.

A specific dataset of classified ground truth points for different trajectories was prepared, and clustering parameters were experimentally set as follows: $MinPts = \frac{1}{7}k$ and $Eps = \frac{1}{3}k$.

Points Of Interest To achieve visual recognition of cultural Points Of Interest, we build a set of specific classifiers. In particular, based on the georeferenced route of the visitor, we retrieve points of high cultural interest querying geolocalized DBpedia for a set of four classes, chosen from main Wikipedia categories of particular interest in cultural heritage (i.e. Buildings and structures by location, Monuments and memorials, Religious architecture, Museums), after which we name our four corresponding classes respectevly Buildings, Monuments, Churches, Museums. In order to retrieve a sufficient number of reliable and up to date image training samples from the web, we extract from Flickr georeferenced images tagged with the corresponding POI for positive samples, while negative samples are randomly chosen from georeferenced images far from the visitor's location over a threshold.

Once collected the training set, Fisher Vectors based on local SIFT features densely sampled (FV) are extracted. This is done in each region of the spatial pyramid, which was set up combining regions in this configuration: 1×1, 2×2 and 3×1 and the FVs of each of these regions are concatenated for each image.

This results in a vectorial representation x of $D = M \times 2G \times R$ dimensions per image, where $M = 80$ is the local feature dimensionality (after PCA), $G = 256$ is number of Gaussians in the mixture and $R = 8$ is the number of pyramid regions. Point of Interest detection is performed on every ten frames extracted from the video.

3 VAEX system: a web tool for egocentric video personalization

The VAEX system is a multi-layer web system for video personalization. Each uploaded user video is processed and automatically annotated on different semantic dimensions which rely on geolocalization and on web leveraged knowledge of the surrounded cultural POI, on camera's wearer behaviour and information about stops and moves.

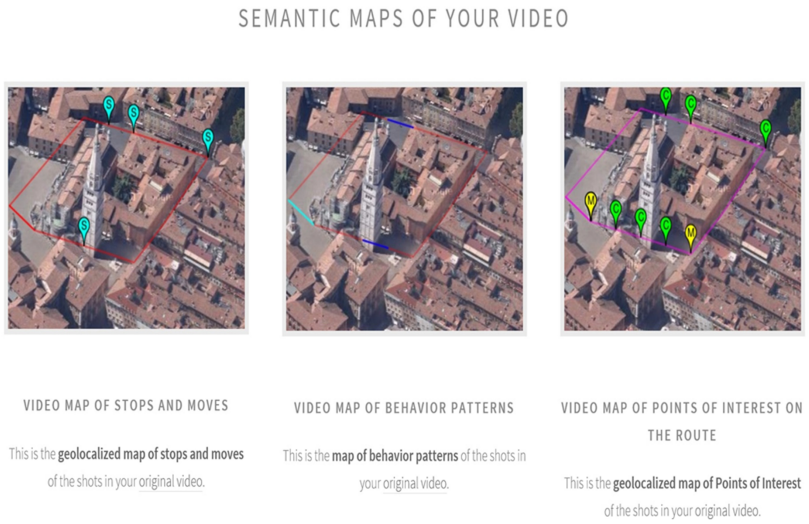


Fig. 1. VAEX Tool: Video semantic browsing. Legend for the Figure 1 on the left: spatial stop shots are marked with teal markers, red continuous line corresponds to move pattern shots. Figure 1 in center: Behavior pattern: blue continuous line = attention; cyan continuous line = wandering, red continuous line = traveling, light green continuous line=changing point of attention (head motion). Figure 1 on the right: light green markers labeled with "C" ="Church" or "Cathedral", yellow markers labeled with "M" = "Monument", red markers labeled as "S" = "Museum" or "Exhibition", cyan markers labeled as "B" = "Palaces" or "Buildings".

The interface shown in Figure 1 allows the user to browse the video according to any of the three semantic level separately by clicking the corresponding image

in the main interface. Selecting a specific semantic dimension the user may easily browse along the shots labeled on that dimension as shown in Figure 2.

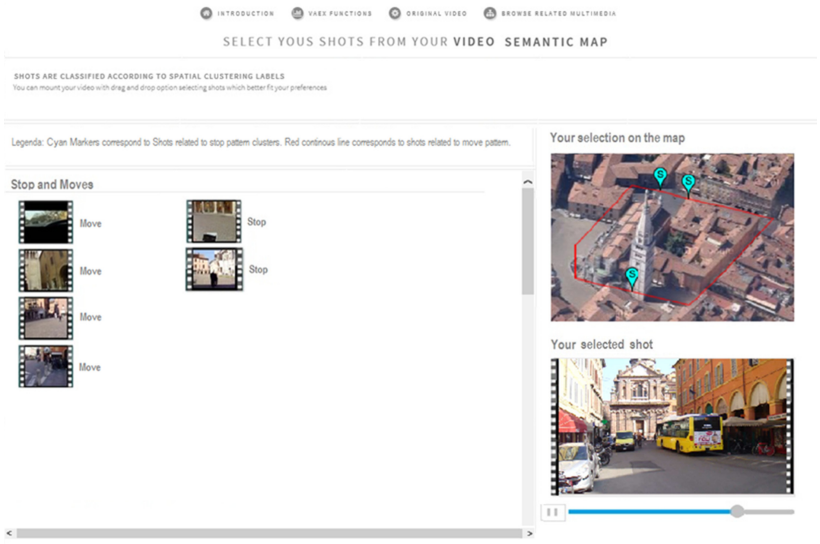


Fig. 2. VAEX Tool: Video semantic browsing user interface to explore video in according to a stops and moves semantic dimension.

The VAEX system includes as the main feature the video personalization interface (see Figure 3) where the semantic dimensions are shown in separated columns and can be crossed together. Shots, labeled with all the recognizable tags within the correspondent semantic category (see for example Modena Cathedral in the POI dimension), may be specifically selected according to the user preferences through the search bar, and drag and dropped in the working timeline to build the personalized stream.

4 Experimental Results

4.1 Behavior pattern detection

To evaluate the performance of the proposed behavior pattern detection, we collected ten videos from head-mounted cameras captured by tourists that spend some time to visit a cultural city. 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. The camera is placed on the tourist's head and captures a 720×576, 24 frames per second RGB image sequence. Granularity of GPS sensor is one second in time and 2 meters linear displacement in space.

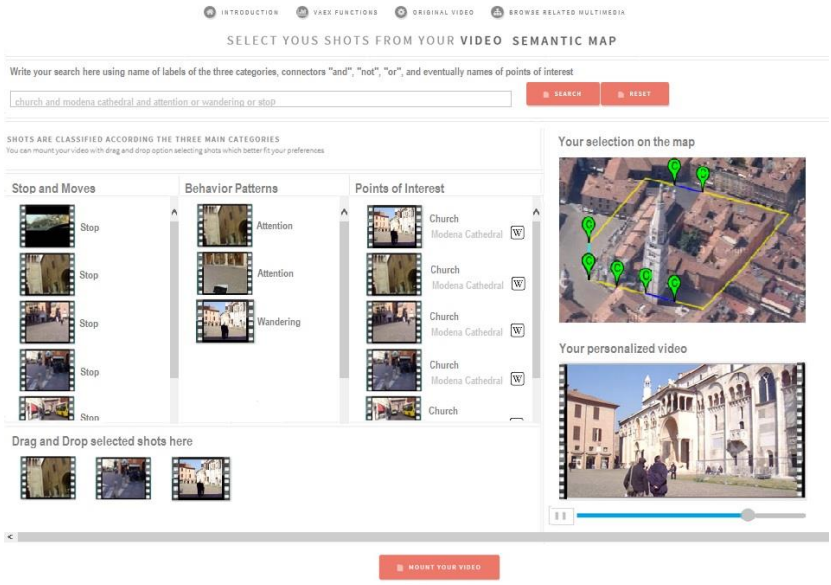


Fig. 3. VAEX Tool: Video personalization general interface with search function.

A subset of 7200 annotated frames is used in order to test our methodology to recognize the high level observer's behaviors: "Attention", "Transit", "Changing point of interest", "Wandering around". First, we examine the effectiveness of our 297-dimension feature vector, based on blurriness, optical flow and acceleration spatial gradient directions weighted over magnitudes and magnitudes, on 3×3 grids with average pooling in time, on low-level motion pattern detection: "Static", "Walking", "High speed", "On wheels", "Head Roll" and "Head Pitch".

In Table 1 we compare our results to a similar descriptor recently proposed by Lu *et al.* [9] (25-dimension feature vector based on blurriness, optical fow directions weighted over magnitudes and magnitudes). The Figure 4 shows the performance of the two techniques per class.

	Lu *et al.* [9]	**Our approach**
Accuracy	62.92	**72.48**

Table 1. Comparison of classification accuracy.

As can be seen from Figure 4, adding feature vectors related to optical flow variations in magnitude and orientation over each of the 3×3 frame sections, with 8-bins quantization, we achieve a better precision as optical flow variation represents local motion thus helps distinguish the special motion of abrupt and

random camera movements, related to "Head motion" and "Head pitch", from significant motion.

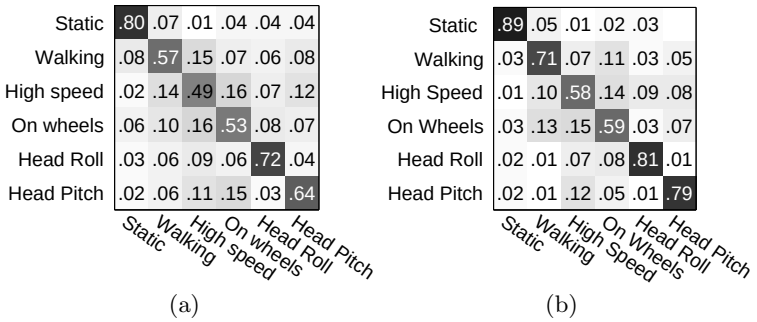
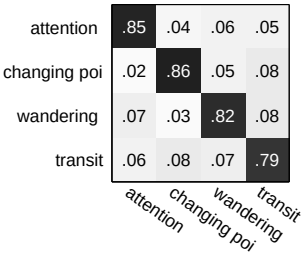


Fig. 4. Classification accuracy using different descriptors: a) feature vector proposed by Lu *et al.* [9]; b) our feature vector.

Finally Figure 5 presents the results obtained applying to the primitive motion classification a Hidden Markov Model, to recognize the behavior patterns and gain classification smoothing, and shows that the accuracy results are quite promising. "Attention" and "Changing POI" obtain a higher performance with respect to the other two classes. This is probably due to the fact that these two last behaviors are characterized by different types of motion caused by the combination of head and body movements and the fast background changes.

Fig. 5. HMM estimated behavior pattern detection accuracy.



4.2 Participants and experiments

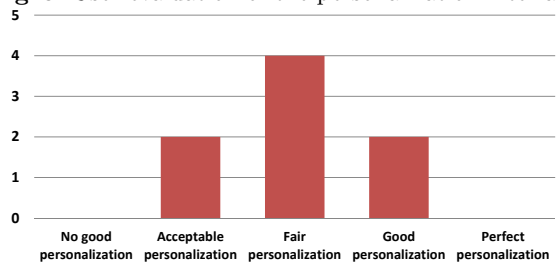
A subjective evaluational test was performed by selecting to participate 8 subjects. Seven participants were undergraduate and postgraduate students and one from technical staff, all of them ranging in age from 18 to 45. They had

no previous knowledge about video personalization or video editing. Participants self-reported that they were familiar with web searches and most common programs for editing of text and images with GUI interface. The main qualifying criterion for a participant in entering the evaluation experiment was to have a strong interest in the fruition of common online video platforms for user generated video and to have a certain familiarity with text and images processors and presentation tools. The subjects were requested to produce their own personalization, working on videos belonging to our egocentric dataset. Since personalization of all 10 movies might have been burdensome for some subjects, they were randomly divided into 2 groups of 4 subjects each.

Subjects were first invited for a twenty minutes tutorial session, in which they were given instructions about the system and shown how to specify preferences to personalize videos.

Finally, a "blind taste test" was performed, in which each group had to evaluate each personalized video by the other group w.r.t the user expressed preferences. We used a Likert scale with a score between 1 and 5, where 1 was "no good personalization" and 5 "perfect personalization" w.r.t user preferences. This test, resulting in Fig. 6, shows that 75% of the evaluations considers the web application a useful and suitable tool for building a short and customized personal video.

Fig. 6. User evaluation of the personalization interface.



5 Conclusions

In this paper we have proposed a video personalization web system designed to support tourists to personalize the egocentric captured videos of their experiences, based on shots automatic classification according to the semantic dimensions of stops and moves, POI detection and behavior pattern. The system supports the user in semantic browsing through the scenes of the video and in selecting and combining easily on the working timeline the relevant shots. The experimental assessments reported in Section 4 exhibit promising results, from the point of view of both results accuracy and usefulness of the personalization web tool.

Bibliography

[1] R. D. Araújo, T. Brant-Ribeiro, R. G. Cattelan, S. A. d. Amo, and H. N. Ferreira. Personalization of interactive digital media in ubiquitous educational environments. In *Proc. of SMC*, 2013.

[2] M. Campanella, R. Leonardi, and P. Migliorati. The future-viewer visual environment for semantic characterization of video sequences. In *Proc. of ICIP*, 2005.

[3] 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.

[4] 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.

[5] G. Farnebäck. Two-frame motion estimation based on polynomial expansion. In *Proc. of the 13th Scandinavian Conference on Image Analysis*, 2003.

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

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

[8] 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.

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

[10] C. Snoek, K. Sande, O. d. Rooij, B. Huurnink, J. Uijlings, M. v. Liempt, M. Bugalhoy, I. Trancosoy, F. Yan, M. Tahir, et al. The mediamill trecvid 2009 semantic video search engine. In *Proc. of TRECVID*, 2009.

[11] Y. Wei, S. M. Bhandarkar, K. Li, and L. Ramaswamy. Video personalization in heterogeneous and resource constrained environments. *Multimedia Systems Journal*, 17(6):523–543, 2011.

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