# Video Summarization in Social Media Based on Users' Geo-location

Klimis Ntalianis and Nikos Mastorakis

***Abstract*—**Tons of information is posted everyday on social networks. This information should be summarized in order to be accessible by users. In this paper a novel scheme is proposed for summarizing video content posted on social networks. Towards this direction a probabilistic framework for estimating the location of each post is designed, based on the associated textual content of the post. Furthermore a CLARANS-based key-frames extraction scheme is considered. Then each user is presented with a location-aware summary. The nearest a video is to the location of a user, the more key-frames are extracted. This paper forms an initial study of location-based summarization of video content posted on social networks and experiments on real world data indicate its promising performance.

***Keywords*—** CLARANS, Fuzzy Representation, Geo-location, Social Media, Twitter, Video Summarization.

## I. Introduction

In the digital era we live there is a massive progress and development of the internet and online world technologies. As a result huge volumes of data are gathered day by day from many different resources and services. This data comes from different online resources and services including Sensor Networks, Cloud Storages, Social Networks etc. These big volumes of data need to be managed and reused so that data analytics are provided. Furthermore although this massive volume of data can be really useful for people and corporations it is problematic as well. They need big storages and their volume makes operations such as processing, search and retrieval real difficult and highly time consuming. One solution to the problem is to summarize big data so they would need less storage and extremely shorter time to get processed and retrieved. The summarized data will then be in a compact form but still an informative version of the entire data.

This paper focuses on video content posted on social networks. The difficulty in video summarization lies in the definition of "important". Which are the important video segments ? Which are the important key-frames ? How can they be extracted ? In the literature several approaches have focused on a certain type of video. For example, the "goal event" is an important moment for a football game. However in generic videos, where the structure, type, actors etc. are not known, it is difficult to define (in term of machine rules) the "important" parts. In this work we give emphasis on the geographical location of each video, by also considering the location of each user.

In particular in this paper we focus on video content, shared by the accounts we follow on Twitter. Towards this direction we aim at providing to users a meaningful, well-organized, interactive and personalized summary of the video content that has appeared on their timelines. In fact, during summarization the geographical location of both the user and the video are also taken into consideration. In this framework one successful way is to summarize video sequences and several works, either key-frame based (i.e. storyboard or static summarization) or sequence based (i.e. video skim) have been proposed in the literature (please see Section II).

On the other hand, content posted on social networks has been used in a variety of applications such as for the ranking of news stories [1], for the profiling of user preferences [2] and even for products' recommendations [3]. However it has not been extensively used for video summarization and the work of Hannon et. al [4] is one of the first to be done towards this direction. In particular time-stamped opinions are utilized for generating soccer video highlights. The introduced PASSEV evaluates two basic summarization approaches. However PASSEV takes as input a video sequence and a collection of time-stamped tweets that are known to refer to the event captured by the video, information that is not available in most cases. Furthermore PASSEV focuses on soccer, excluding many other popular video content categories. Finally PASSEV does not consider geographical information of users and/or videos, during the production of summaries.

In this paper we want to examine the capability of automatically producing a meaningful summary of generic videos posted on social networks. In this work the summarization process is guided by geographical information (location) of content and users. In particular, one of the main concepts of this scheme has to do with the geo-location. We believe that there are several persons that prefer to see videos, the content of which refers to regions near these persons. For example if someone lives in New York, he/she may prefer to see videos containing events that happen in New York. However this concept does not exclude videos that refer to locations far away from the users. It only reduces the number

K. Ntalianis is with the Athens University of Applied Sciences, Department of Marketing, Agiou Spyridonos, Egaleo, Athens, GREECE (e-mail: kntal@ teiath.gr).

N. Mastorakis is with the Technical University of Sofia, Department of Industrial Engineering, Sofia, BULGARIA, (e-mail: mastor@tu-sofia.bg).

of key-frames according to distance. In any case the user can retrieve the whole video if he/she is interested in its (limited for distant videos) key-frames.

In order to extract more key-frames for nearby videos, initially the location of each video is estimated based on a Kernel Density Estimation (KDE) approach, which is properly adapted for geo-location analysis. Next the distance of each video from the location of each user is used to tune the CLARANS algorithm, which extracts the necessary numbers of medoids (key-frames in our case). We believe that this paper serves as one initial study for providing large scale geo-location based summarization of social media content.

The rest of the paper is organized as follows: Section II describes previous works. In Section III the geo-location estimation algorithm based on text analysis is analyzed. A detailed description of the video summarization method is provided in Section IV. Experiments on real life data are presented in Section V. Finally Section VI concludes this paper providing also directions for future research.

## II. Previous Work

Video summaries can be created in various forms, but the most well known are the forms of key-frames or video skims. Sets of key-frames (also called static storyboards), represent the main content of a video sequence using a group of salient frames. In video skimming a video clip is extracted with a much shorter duration than the original video. The main aim of video summarization is to algorithmically engineer computers to conceive specific multimedia data by giving the computers the ability to interpret multimedia data in the same manner as human perception. On the other hand video summaries can be generated manually or automatically. However manual summarization is not feasible if we think of the huge volumes of video data and the limited manpower. Thus, the development of fully automated video analysis algorithms is a very important and challenging research area.

Towards this direction, many video summarization approaches have been proposed in the literature. In one of the early tries [5], video frames were represented as a trajectory curve and a generalized version of the planar curve splitting algorithm was recursively applied to simplify the curve. In [6] key-frame selection was formulated as a temporal rate-distortion MINMAX optimization problem and dynamic programming was employed to obtain the optimal solution. In [7] a set of modeling methods for visual and aural attentions were proposed. In [8] a graph connectivity technique and a dominant set clustering method were combined for automatic keyframes' selection. In [9] semantic video summarization is performed by trying to answer the «who», «what», «where», and «when» questions.

Characters that appear in movies are also used as domain knowledge. For these domains, various types of metadata help to generate video summaries [10]-[12]. Egocentric videos are another interesting example, for which a video summarization approach using a certain set of predefined objects as a type of domain knowledge has been proposed [13]. Potapov et al. [14] proposed to summarize a video focusing on a specific event and used an event classifier's confidence score as the importance of a video segment. Yang et al. [15] proposed to utilize an auto-encoder, in which its encoder converts an input video's features into a more compact one, and the decoder then reconstructs the input. Additionally the diversity of segments included in a video summary is an important criterion and many approaches use various definitions of the diversity [16]-[18]. Canonical views of visual concepts can be an indicator of important video segments, and several existing work uses this intuition for generating a video summary [19]-[21].

The supervised approaches of [22], [23] learn to combine multiple hand-crafted criteria so that the summaries are consistent with ground truth. In [24] an approach to obtain a representative and diverse summary by clustering videos into events and selecting the best frame per event is proposed. In [25] an approach for query-adaptive video summarization using DPPs (Determinantal Point Processes) is proposed. The method is limited to a small, fixed set of concepts such as "car" or "flower". In [26], a summarization technique was specifically proposed for producing on-the-fly video storyboards. This method produces still and moving storyboards and it is based on a fast clustering algorithm, which selects the most descriptive visual frames based on the HSV frame color distribution.

Even though interesting, most of the aforementioned works cannot be straightforwardly applied to social media content, while they do not consider geo-location during the summarization phase.

## III. Geo-location Estimation Based on Text Analysis

Standard approaches on geo-location estimation based on textual analysis aim at finding location-relevant terms. One of the most characteristic ones is the $x^2$ term selection. This method uses the $x^2$ statistic to assess to what extent there is a statistically significant difference between the actual number of occurrences of a term $t$ in documents of class $c$ and the number of occurrences that we would expect to see if the probability of seeing $t$ did not depend on whether the corresponding document is in class $c$. Let $O_{tc}$ be the number of times term $t$ occurs in a document of class $c$. Let $O_{t\bar{c}}$ be the number of times $t$ occurs in documents outside $c$, $O_{\bar{t}c}$ the number of occurrences of terms other than $t$ in documents of class $c$, and $O_{\bar{t}\bar{c}}$ the number of occurrences of terms other than $t$ in documents outside class $c$. Moreover, $E_{tc}$ is the expected frequency of term $t$ in class $c$ (similar for $E_{\bar{t}c}$, $E_{t\bar{c}}$ and $E_{\bar{t}\bar{c}}$):

$$E_{tc} = \frac{\sum_{c} O_{tc}}{\sum_{c'} \sum_{t'} O_{t'c'}} \cdot \sum_{t'} O_{t'c} \tag{1}$$

$$E_{\bar{t}c} = \frac{\sum_c \sum_{t' \neq t} O_{t'c}}{\sum_{c'} \sum_{t'} O_{t'c'}} \cdot \sum_{t'} O_{t'c} \tag{2}$$

$$E_{t\bar{c}} = \frac{\sum_c O_{tc}}{\sum_{c'} \sum_{t'} O_{t'c'}} \cdot \sum_{c' \neq c} \sum_{t'} O_{t'c'} \tag{3}$$

$$E_{\bar{t}\bar{c}} = \frac{\sum_c \sum_{t' \neq t} O_{t'c}}{\sum_{c'} \sum_{t'} O_{t'c'}} \cdot \sum_{c' \neq c} \sum_{t'} \; O_{t'c'} \tag{4}$$

The most discriminative terms of class $c$ are then chosen as those that maximize the $x^2$ statistic:

$$\begin{aligned} x^2(t,c) = {} & \frac{(O_{tc} - E_{tc})^2}{E_{tc}} + \frac{(O_{\bar{t}c} - E_{\bar{t}c})^2}{E_{\bar{t}c}} + \\ & + \frac{(O_{t\bar{c}} - E_{t\bar{c}})^2}{E_{t\bar{c}}} + \frac{(O_{\bar{t}\bar{c}} - E_{\bar{t}\bar{c}})^2}{E_{\bar{t}\bar{c}}} \end{aligned} \tag{5}$$

However in geotagging there are not natural categories where we can evaluate the $x^2$ statistic. For this reason a Kernel Density Estimation (KDE) approach is adopted and extended for geo-location estimation [27]. In particular let $A$ be a grid of size 512×512. Each tag $t$ is then associated with a probability distribution $p(A \mid t)$ of locations, where locations are the cells of the grid $A$. Then a standard KDE is computed by:

$$q(A|t) = \frac{1}{|A|} \sum_{a \in A} K \frac{(t - t_a)}{\theta}, \quad t \in \Re \tag{6}$$

where $\theta$ is a bandwidth parameter in degrees latitude/longitude and K is the kernel. From the set of all locations of the photos belonging to the training set, a background distribution $p(A)$ is estimated using KDE. In this case the KDE process will provide the distributions $p_{KDE}(A \mid t)$ and $p_{KDE}(A)$. The more the probability distribution $p(A \mid t)$ is centered around a few peaks, the lower the entropy of that distribution will be and the more desirable tag $t$ is for geo-location estimation. However, when estimating $p(A \mid t)$ based on KDE, the total number of occurrences of tag $t$ is not taken into consideration. To cope with this, a further smooth of $p(A \mid t)$ is accomplished by using Bayesian smoothing with Dirichlet priors:

$$p_{Dir}(a|t) = \frac{p_{KDE}(a|t) \cdot N_t + \mu \cdot p_{KDE}(a)}{N_t + \mu} \tag{7}$$

where $N_t$ is the total number of occurrences of tag $t$ and the parameter $\mu \in [0, +\infty)$ controls the number of samples that should be observed in order to abandon the idea that occurrences of $t$ follow the general distribution. For lower values of $\mu$ more rare tags will be selected, while for very large $\mu$, $p_{Dir}(A \mid t)$ will tend to $p_{KDE}(A)$. Using a uniform prior:

$$p_{uni}(a|t) = \frac{p_{KDE}(a|t) \cdot N_t + \frac{\mu}{|A|}}{N_t + \mu} \tag{8}$$

After smoothing entropy can be used for tag ranking:

$$s_{Dir}^{ent}(t) = H_{Dir}(A|t) = -\sum_{a \in A} p_{Dir}(a|t) \cdot \log(p_{Dir}(a|t)) \tag{9}$$

$$s_{uni}^{ent}(t) = H_{uni}(A|t) = -\sum_{a \in A} p_{uni}(a|t) \cdot \log(p_{uni}(a|t)) \tag{10}$$

The idea of (9)-(10) is that useful terms occur at a few selected locations. By reversing this idea we assume that terms are location-relevant to the extent that the distribution of their occurrences diverges from the background distribution $p_{KDE}(A)$. Using the Kullback-Leibler divergence between $p_{KDE}(A)$ and $p_{Dir}(A \mid t)$, to quantify this idea, we have:

$$\begin{aligned} s^{kl}(t) \Big| &= D_{KL}(p_{Dir}(A|t) \| p_{KDE}(A) = \\ &= \sum_{a \in A} p_{Dir}(a|t) \cdot \log(\frac{p_{Dir}(a|t)}{p_{KDE}(a)}) \end{aligned} \tag{11}$$

Equation (11) smoothes out any artifacts from the training set. To this effect, a goodness-of-fit test can be used to assess with which degree of confidence, the null hypothesis can be rejected that the occurrences of tag $t$ have been sampled from $p_{KDE}(A)$. Using the $x^2$ test we take the score:

$$s^{x^2}(t) = \sum_{a \in A} \frac{(O_{ta} - p_{KDE}(a) \cdot N_t)^2}{p_{KDE}(a) \cdot N_t} \tag{12}$$

where $O_{ta}$ is the number of occurrences of tag $t$ in grid cell $a$, and $N_t$ the total number of occurrences of $t$.

## IV. Geo-location Based Video Summarization

According to [28] the majority of Twitter users (82%) watch video content on Twitter and most watch on a hand-held screen. A staggering 90% of Twitter video views happen on a mobile device. But Twitter users are also 1.9 times more likely to have uploaded a video online (anywhere) than the average U.S. internet user. Many would like to see more of breaking news (64%), clips from live sports shows (54%) and clips from TV shows (50%). And Twitter users say they want to see more videos from three top sources: celebrities (45%), other users (40%) and brands (37%). However there is also a need to see videos that are near one's location. This is called location-based video consumption. By considering that the video content that is posted on Twitter everyday amounts to thousands of hours, algorithms are needed to summarize all these hours to few hours.

Towards this direction, in this paper a key-frames extraction algorithm is proposed, focusing on Youtube links posted on Twitter. One of the innovations of the proposed scheme is that it takes into consideration the geo-location of each user. For example a user that is located in Athens should receive a different video summary compared to a user that is located in Thessaloniki. In particular the text of each tweet is analyzed based on the method described in Section III. Then a geo-location is estimated for each tweet, or in other words the actual location which the tweet refers to (e.g. "Fire in the centre of Athens" => the tweet refers to Athens). In this case the posted video link (youtube video) is associated to the location of its related tweet. Next only users that have defined their location field are considered. In this framework, videos that are assumed to refer to a location near the location of a specific user provide more key-frames to the summary, while videos that are far away provide less frames to the summary. As a result a location-based summary is provided to each user.

More specifically, initially each video sequence (that is within a location range of *LR* kilometres for a specific user) is segmented into shots and a fuzzy feature vector is formulated for each frame [29]. In particular let us assume that an examined frame consists of *K* segments (color, motion). Each feature $s_i$, $i$=1, …, *K*, of the *i*th segment can be classified to *Q* classes using *Q* membership functions $\mu_n(s)$, $n$ = 1, 2, …, *Q*. Then, the degree of membership of all *K* segments of the respective frame to the *n*th class is calculated through the fuzzy histogram, say $H(n)$:

$$H(n) = \frac{1}{K}\sum_{i=1}^{K}\mu_n(s_i), \quad n = 1,2,...,Q \tag{13}$$

In this paper we assume that $K^c$ color and $K^m$ motion segments are extracted for each frame. Then for each color segment $S_i^c$, $i$=1, …, $K^c$, an $L^c \times 1$ vector $\boldsymbol{s}_i^c$ is formed, while for each motion segment $S_i^m$, $i$=1, …, $K^m$, an $L^m \times 1$ vector $\boldsymbol{s}_i^m$ is formed as:

$$\mathbf{s}_i^c = [\mathbf{c}^T(S_i^c)\mathbf{l}^T(S_i^c)a(S_i^c)]^T \tag{14}$$

$$\mathbf{s}_i^m = [\boldsymbol{\upsilon}^T(S_i^m)\mathbf{l}^T(S_i^m)a(S_i^m)]^T \tag{15}$$

where $\alpha$ denotes the size of the color or motion segment and $\boldsymbol{l}$ is a 2 × 1 vector indicating the horizontal and vertical locations of the segment center; the 3 × 1 vector **c** includes the average values of the three color components of the respective color segment, while the 2 × 1 vector $\boldsymbol{\upsilon}$ includes the average motion vector of the motion segment.

For notational simplicity, superscripts *c* and *m* will be omitted in the sequel and each color or motion segment will be denoted as $S_i$ and will be described by vector $\mathbf{s}_i$. Thus:

$$\mathbf{s}_i = [s_{i,1}\ \ s_{i,2} \cdots s_{i,L}]^T \tag{16}$$

is a vector containing all properties extracted from the *i*th segment $S_i$. Each element $s_{i,j}$, $j$=1,2, …, *L* of vector $\boldsymbol{s}_i$ is then partitioned into *Q* regions by means of *Q* membership functions $\mu_{n_j}(s_{i,j})$, $n_j = 1,2,...,Q$. Now $\mu_{n_j}(s_{i,j})$ denotes the degree of membership of $s_{i,j}$ to the $n_j$th class. Then the product of $\mu_{n_j}(s_{i,j})$ over all $s_{i,j}$ of $\boldsymbol{s}_i$ defines the degree of membership of vector $\boldsymbol{s}_i$ to the L-dimensional class $\boldsymbol{n} = [n_1 n_2 ... n_L]^T$, the elements of which express the class to which the elements of $\boldsymbol{s}_i$ belong.

$$\mu_n(\mathbf{s}_i) = \prod_{j=1}^{L}\mu_{n_j}(s_{i,j}) \tag{17}$$

Gathering all segments of a frame, a multidimensional fuzzy histogram is created:

$$H(\mathbf{n}) = \frac{1}{K}\sum_{i=1}^{K}\mu_n(\mathbf{s}_i) = \frac{1}{K}\sum_{i=1}^{K}\prod_{j=1}^{L}\mu_{n_j}(s_{i,j}) \tag{18}$$

$H(\boldsymbol{n})$ can be viewed as a degree of membership of a whole frame to class $\boldsymbol{n}$. A frame feature vector $\boldsymbol{f}$ is then formed by gathering all values of $H(\boldsymbol{n})$ for all classes $\boldsymbol{n}$, resulting in a vector of $Q^L$ elements: $\boldsymbol{f} = [f_1 f_2 \cdots f_{Q^L}]^T$. In particular vector $\boldsymbol{f}$ is constructed from $H(\boldsymbol{n})$ using an index function $z(\boldsymbol{n})$ which maps the class $\boldsymbol{n}$ to an integer between 1 and $Q^L$,

$$z(\mathbf{n}) = 1 + \sum_{j=1}^{L}(n_j - 1)Q^{L-j} \tag{19}$$

Then the elements $f_i$, $i$=1, …, $Q^L$, of $\boldsymbol{f}$ are calculated as $f_{z(n)} = H(\boldsymbol{n})$ for all classes $\boldsymbol{n}$. In fact, since the above analysis was based on features $\boldsymbol{s}_i^c$ and $\boldsymbol{s}_i^m$ of color $S_i^c$ and motion $S_i^m$ segments, respectively, two feature vectors will be calculated: a color feature vector $\boldsymbol{f}^c$ and a motion feature vector $\boldsymbol{f}^m$. Thus the total feature vector of an image is:

$$\mathbf{f} = [(\mathbf{f}^c)^T(\mathbf{f}^m)^T]^T \tag{20}$$

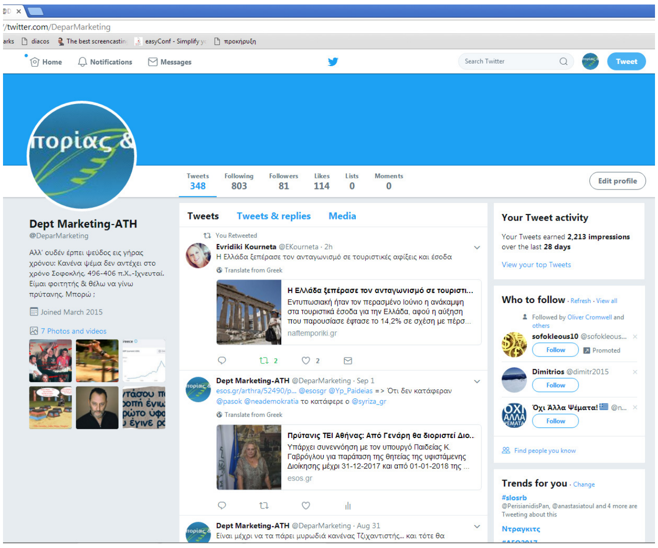
Next, spatial clusters are created for each shot, the medoids of which are selected as key-frames. In our work, the CLARANS algorithm [30] has been adopted for spatial clustering due to its low complexity, scalability and quality of results. In particular let us consider that key-frames should be extracted from each shot with duration > $T_s$. According to CLARANS, the process of finding *k* medoids among *n* points of a space, can be viewed abstractly as searching through a certain graph. In such a graph, denoted by $G_{n,k}$, each node represents a set of *k* points $\{MD_{m,1}, …, MD_{m,k}\}$ of an *M*-dimensional space, indicating that $MD_{m,1}, …, MD_{m,k}$ are the selected medoids. Two nodes $ND_1$ and $ND_2$ are considered as neighbors if their sets differ by only one point. More formally, for

$$ND_1 = \{MD_{m,1}, …, MD_{m,k}\}$$
$$\text{and} \tag{21}$$
$$ND_2 = \{MD_{w,1}, …, MD_{w,k}\}, |ND_1 \cap ND_2| = k\text{-}1$$

where || is the cardinality of the intersection. Each node has $k(n - k)$ neighbors. Since a node represents a collection of *k* medoids, each node corresponds to a clustering and can be assigned a cost (e.g. the total dissimilarity between every point and the medoid of its cluster). Here the differential of [30] is used for cost estimation of each node. The CLARANS algorithm has two parameters: *maxneighbor* and *numlocal*. The first determines the maximum number of neighbors that are examined in each iteration, while the second determines the number of local minima that should be searched. In this paper the values of *maxneighbor* and *numlocal* together with the number of *k* medoids are controlled by the geo-location of each video. In particular the nearest the location of a video is to a specific user, the more the extracted key-frames (*k* medoids) and the finer the detail of the search (*maxneighbor* and *numlocal* parameters). In the experimental results section more details are provided.

## V. Experimental Results

In order to evaluate the performance of the proposed geo-location based video summarization scheme, we have carried out different experiments. Initially we have recorded and archived Twitter's newsfeed information of 81 followers of the account of the Department of Marketing of the Athens University of Applied Sciences (Figure 1) (https://twitter.com/DeparMarketing) for a period of 60 days (16 July 2017 – 14 September 2017), using the twitteR package of the R language. During archiving we have discarded all other posts except of Youtube videos and have kept their exact order as presented on the walls of the 81 followers. In total 1,223 videos have been gathered together with their associated text. The average video duration was 174 seconds, providing a total of 5,320,057 frames.



| Tweets | Following | Followers | Likes | Lists | Moments |
|---|---|---|---|---|---|
| 348 | 803 | 81 | 114 | 0 | 0 |

Dept Marketing-ATH
@DeparMarketing

Αλλ' ουδέν έρπει ψεύδος εις γήρας χρόνου: Κανένα ψέμα δεν αντέχει στο χρόνο Σοφοκλής, 496-406 π.Χ.,-Ίχνευταί. Είμαι φοιτητής & θέλω να γίνω πρύτανης. Μπορώ ;

Joined March 2015

7 Photos and videos

You Retweeted

Evridiki Kourneta @EKourneta · 2h
Η Ελλάδα ξεπέρασε τον ανταγωνισμό σε τουριστικές αφίξεις και έσοδα
Translate from Greek

**Η Ελλάδα ξεπέρασε τον ανταγωνισμό σε τουριστι...**
Εντυπωσιακή ήταν τον περασμένο Ιούνιο η ανάκαμψη στα τουριστικά έσοδα για την Ελλάδα, αφού η αύξηση που παρουσίασε έφτασε το 14,2% σε σχέση με πέρσ...
naftemporiki.gr

2 2

Dept Marketing-ATH @DeparMarketing · Sep 1
esos.gr/arthra/52490/p... @esosgr @Yp_Paideias => Ότι δεν κατάφεραν @pasok @neademokratia το κατάφερε ο @syriza_gr
Translate from Greek

**Πρύτανις ΤΕΙ Αθήνας: Από Γενάρη θα διοριστεί Διο..**
Υπάρχει συνεννόηση με τον υπουργό Παιδείας Κ. Γαβρόγλου για παράταση της θητείας της υφιστάμενης Διοίκησης μέχρι 31-12-2017 και από 01-01-2018 της ...
esos.gr

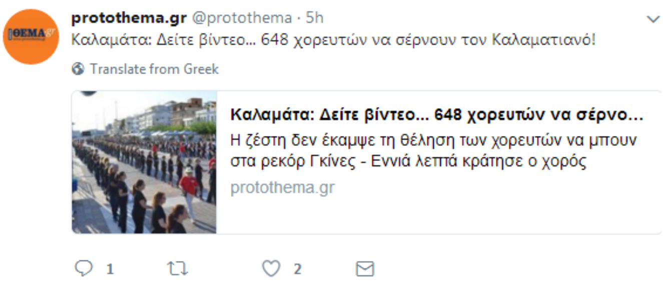
Dept Marketing-ATH @DeparMarketing · Aug 31
Είναι μέχρι να τα πάρει μυρωδιά κανένας Τζιχαντιστής... και τότε θα

**Your Tweet activity**
Your Tweets earned **2,213 impressions** over the last **28 days**
View your top Tweets

**Who to follow** · Refresh · View all
Followed by Oliver Cromwell and others
sofokleous10 @sofokleous... Promoted
Dimitrios @dimitr2015
Όχι Άλλα Ψέματα! @n...

**Trends for you** · Change
#slosrb
@PerisianidisPan, @anastasiatoul and 4 more are Tweeting about this
Ντραγκιτς

Fig. 1: Overview of the twitter account of the Department of Marketing


**protothema.gr** @protothema · 5h
Καλαμάτα: Δείτε βίντεο... 648 χορευτών να σέρνουν τον Καλαματιανό!
Translate from Greek

**Καλαμάτα: Δείτε βίντεο... 648 χορευτών να σέρνο...**
Η ζέστη δεν έκαμψε τη θέληση των χορευτών να μπουν στα ρεκόρ Γκίνες - Εννιά λεπτά κράτησε ο χορός
protothema.gr

1 2

Fig. 2: An example post and its associated text

Next the geo-location estimation module was activated, which classified all videos to geographical locations, according to their associated text. For example in case of Figure 2, the associated text says: "Καλαμάτα: Δείτε βίντεο... 648 χορευτών να σέρνουν τον Καλαματιανό!" (Kalamata: Watch the video … 648 dancers, dancing Kalamatiano). Among the 1,223 videos, 10.96% of the videos (134) were classified to the location "Greece" while the rest 89.04% referred to 18 more countries (USA, UK, Germany, Netherlands, Belgium, Spain, France etc). This is expected since most of the followers are from Greece and they follow several Greek persons, media, companies etc.

Afterwards the video summarization module was activated for the archived ordered content. Firstly videos were downloaded. In the next step each video was segmented into shots and $T_s$ was set equal to 3 sec, so that unperceived visual content is discarded. Additionally $L^c$ = 6, $L^m$ = 5 and $Q$ = 3 (triangular membership function). Then for each frame of the remaining shots, $\boldsymbol{f} = [(\boldsymbol{f}^c)^T (\boldsymbol{f}^m)^T]^T$ was estimated. Summaries were created for each user in a per 10-days period basis. A threshold of 0.1% was also set to the CLARANS algorithm so that it presents on average only 0.1% of the posted video content (or about 22 frames per average duration video). Since CLARANS took into consideration the geo-locations of videos and users, more key-frames were selected from videos of nearby geo-locations (to the location of a user) than from videos of distant geo-location. In particular the abstraction threshold fluctuated between 0.37% (nearest video) and 0.046% (remotest video), eventually providing the average 0.1%. To achieve these results *k* (number of medoids → key frames) took values in the interval [1 22], *maxneighbor* was set equal to 3.75% [30] of the total number of frames in a shot, while *numlocal* was estimated for each shot according to its dispersion and took values in the interval [1 4]. To visualize the bandwidth gain, we define the Information Reduction coefficient as:

$$\text{IR} = \frac{\text{\# of frames transmitted by the traditional approach}}{\text{\# of frames transmitted by the proposed method}} \qquad (22)$$

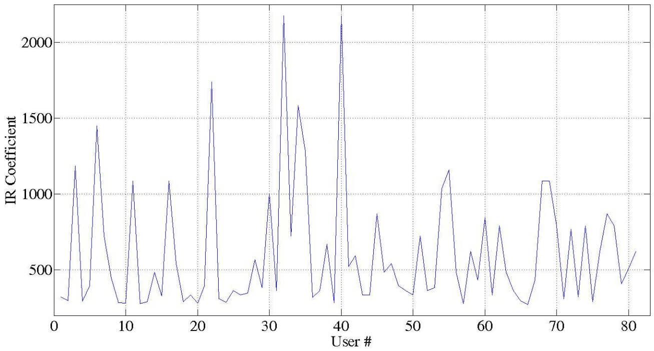


Fig. 3: IR coefficient for the 6th 10-days summary of all users

IR represents a kind of compression ratio. In Figure 3 the IR coefficient for all users is presented, regarding the 6th 10-days summary, where 214 videos were presented on their timelines. As it can be observed, compression ratios fluctuate between 271.88 to 2175 times.

Furthermore Figure 4 illustrates the shot of Fig. 2 which is used to demonstrate the performance of the key-frames extraction algorithm. The shot consists of 495 frames. One out of every 25 frames is depicted, resulting in 20 thumbnails. Figure 3 presents the two extracted keyframes of this shot obtained from the CLARANS algorithm. These two key-frames provide a summary of the overall view (right and left side of the road). Furthermore the provided key-frames of all videos were also connected to the real video segment (by using Youtube's video annotations), so that an interested user could go to the time instance of the frame of interest and watch more parts of the clip.

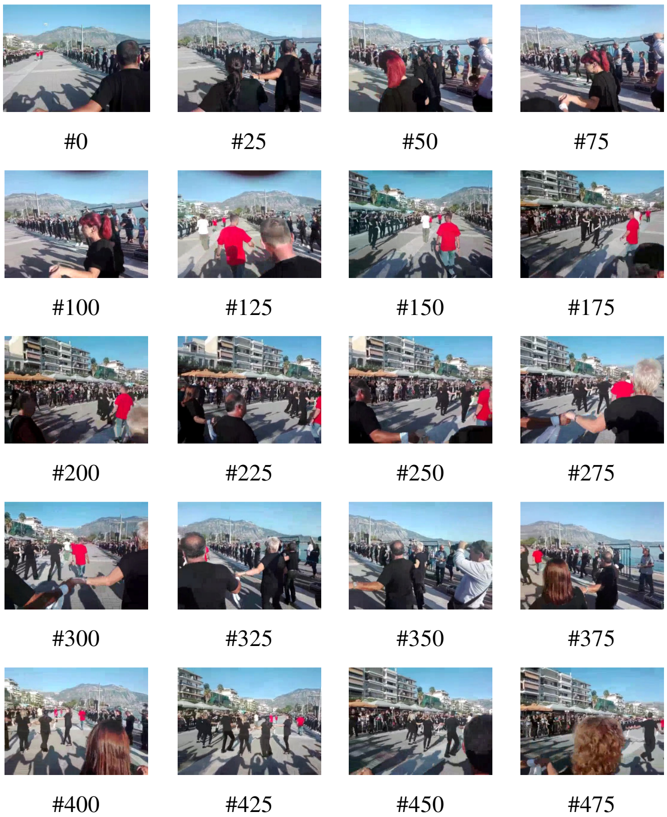

#0 #25 #50 #75
#100 #125 #150 #175
#200 #225 #250 #275
#300 #325 #350 #375
#400 #425 #450 #475

Fig. 4: The video of Fig. 2 consisting of 495 frames, shown with one frame out of every 25

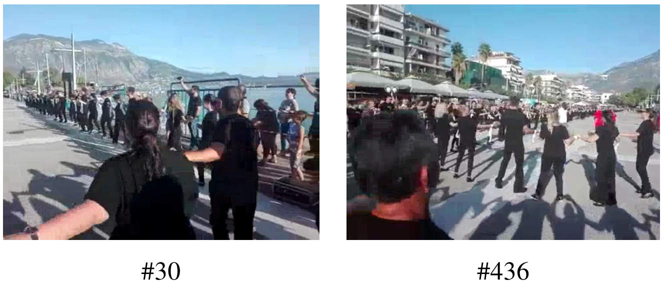

#30 #436

Fig. 5: Two key-frames selected by CLARANS

Finally we have also contacted an experiment to test user satisfaction. Towards this direction we have provided to our 81 users: (a) the summarized videos (10-days period) according to their original order on the users' timelines (b) the summarized videos (10-days period) in a random order and (c) the summarized videos (10-days period) in order of distance (proposed location-based approach). Then each user was asked to express a viewing preference among the three different summaries (or "among the three summaries, which one do you prefer"). Of course summaries did not have any hint about the producing algorithm and the presentation sequence changed so that not to affect the experiment. In total 486 preference sets of the three aforementioned approaches have been collected (six 10-day summaries per user). Results are provided in Figure 6, where the original received 152 preferences (~ 31%), the proposed 196 (~41%) and the random 138 (~28%). As it can be observed differences are not very large. However there is a tendency of preferring geo-location based summarization. More specifically, among the 81 users 21 showed this tendency more explicitly (they mostly cared about events near them and 87% of their preferences were on the proposed scheme). However the other users did not show any specific tendency. This is possibly due to the fact that summaries are just key-frames without containing any audio.

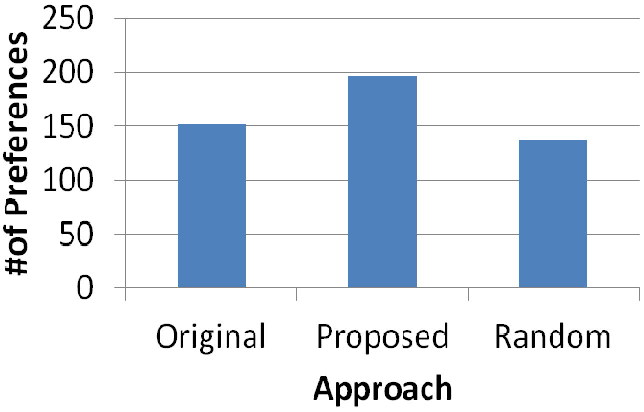


Fig. 6: Users' viewing preferences among three approaches

## VI. Conclusion

Work-in-progress was presented in this paper, focusing on location-based video summarization. In particular a video summarization algorithm was proposed that takes into consideration the geo-location of each video and each user to guide the spatial clustering CLARANS algorithm. Our initial results are very promising since information compression of up to 2175 times is achieved, while users are generally satisfied with the quality of results.

In the future a much larger experimentation phase should be set up with more users, more locations and much more content. One of the main challenges is the estimation of location, especially in cases where the associated text does not contain any location-clarifying keywords. Additionally it would also be interesting to allow users to interact with the summarization module, so that they can arrange the duration, geo-location and composition of summaries.

## Acknowledgment

This research was supported by the World Scientific and Engineering Academy and Society (wseas.org). The authors would like to thank all followers of the account of the Department of Marketing at Twitter.

## References

[1] O. Phelan, K. McCarthy, and B. Smyth, “Using twitter to recommend real-time topical news,” In Proc. of the *3rd ACM conference on Recommender systems*, p.p. 385–388, New York, USA, 2009.

[2] J. Hannon, M. Bennett, and B. Smyth “Recommending twitter users to follow using content and collaborative filtering approaches,” *4th ACM conference on Recommender systems*, p.p. 199–206, New York, USA, 2010.

[3] S. G. Esparza, M. P. O’Mahony, and B. Smyth, “On the real-time web as a source of recommendation knowledge,” In Proc. of the *4th ACM conference on Recommender systems*, p.p. 199–206, New York, USA, 2010.

[4] J. Hannon, K. McCarthy, J. Lynch, and B. Smyth “Personalized and Automatic Social Summarization of Events in Video,” In *Proc. of IUI*, Palo Alto, California, USA, 2011.

[5] D. F. DeMenthon, V. Kobla, and D. Doermann, “Video summarization by curve simplification,” in *ACM International Conference on Multimedia*, 1998.

[6] L. Zhu, G. Schuster, and A. Katsaggelos, “Minmax optimal video summarization,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, pp. 1245–1256, 2005.

[7] Y.-F. Ma, X.-S. Hua, L. Lu, and H.-J. Zhang, “A generic framework of user attention model and its application in video summarization,” *IEEE Trans. on Multimedia*, vol. 7, pp. 907 – 919, 2005.

[8] D. Besiris, A. Makedonas, G. Economou, and S. Fotopoulos, “Combining graph connectivity & dominant set clustering for video summarization,” *Multimedia Tools and Applications*, vol. 44, pp. 161–186, 2009.

[9] B.-W. Chen, J.-C. Wang, and J.-F. Wang, “A novel video summarization based on mining the story-structure and semantic relations among concept entities,” *IEEE Trans. on Multimedia*, vol. 11, pp. 295–312, 2009.

[10] N. Babaguchi, Y. Kawai, T. Ogura, and T. Kitahashi, “Personalized abstraction of broadcasted American football video by highlight selection,” *IEEE Trans. Multimedia*, p.p. 575–586, 2004.

[11] J. Sang, and C. Xu, “Character-based movie summarization,” In Proc. *ACM Int. Conf. Multimedia*, p.p. 855–858, 2010.

[12] G. Evangelopoulos, A. Zlatintsi, A. Potamianos, P. Maragos, K. Rapantzikos, G. Skoumas, and Y. Avrithis, “Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention,” *IEEE Trans. Multimedia*, p.p. 1553–1568, 2013.

[13] Z. Lu and K. Grauman, “Story-driven summarization for egocentric video,” In Proc. *IEEE Computer Society Conf. Computer Vision and Pattern Recognition*, p.p. 2714–2721, 2013.

[14] D. Potapov, M. Douze, Z. Harchaoui, and C. Schmid, “Category-specific video summarization,” In Proc. *European Conf. Computer Vision,* p.p. 540–555, 2014.

[15] H. Yang, B. Wang, S. Lin, D. Wipf, M. Guo, and B. Guo, “Unsupervised extraction of video highlights via robust recurrent auto-encoders,” In Proc. *IEEE Int. Conf. Computer Vision*, p.p. 4633–4641, 2015.

[16] J. Xu, L. Mukherjee, Y. Li, J. Warner, J. M. Rehg, and V. Singh, “Gaze-enabled egocentric video summarization via constrained submodular maximization,” In Proc. *IEEE Computer Society Conf. Computer Vision and Pattern Recognition*, p.p. 2235–2244, 2015.

[17] S. Tschiatschek, R. K. Iyer, H. Wei, and J. A. Bilmes, “Learning mixtures of submodular functions for image collection summarization,” In Proc. *Advances in Neural Information Processing Systems*, p.p. 1413–1421, 2014.

[18] B. Gong, W. L. Chao, K. Grauman, and F. Sha, “Diverse sequential subset selection for supervised video summarization,” In Proc. *Advances in Neural Information Processing Systems,* p.p. 2069–2077, 2014.

[19] Y. Song, J. Vallmitjana, A. Stent, and A. Jaimes, “TVSum: Summarizing web videos using titles,” In Proc. *IEEE Computer Society Conf. Computer Vision and Pattern Recognition,* p.p. 5179–5187, 2015

[20] A. Khosla, E. Hamid, C. J. Lin, and N. Sundaresan, “Large-scale video summarization using web-image priors,” In Proc. *IEEE Computer Society Conf. Computer Vision and Pattern Recognition*, p.p. 2698–2705, 2013.

[21] W.S. Chu, and A. Jaimes, “Video co-summarization: Video summarization by visual co-occurrence,” In Proc. *IEEE Computer Society Conf. Computer Vision and Pattern Recognition,* p.p. 3584–3592, 2015.

[22] M. Gygli, H. Grabner, L. Van Gool, “Video summarization by learning submodular mixtures of objectives,” In *CVPR*, 2015.

[23] M. Gygli, H. Grabner, H. Riemenschneider, and L. Van Gool, “Creating summaries from user videos,” In *ECCV*, 2014.

[24] S. E. Fontes de Avila, A. P. B. Lopes, A. da Luz, and A. de Albuquerque Araujo, “VSUMM: A mechanism designed to produce static video summaries and a novel evaluation method,” *Pattern Recognition Letters,* Vol. 32, No. 1, 2011.

[25] A. Sharghi, B. Gong, and M. Shah, “Query-Focused Extractive Video Summarization,” CoRR abs/1607.05177, 2016.

[26] M. Furini, F. Geraci, M. Montangero, and M. Pellegrini, “STIMO: STIll and Moving video storyboard for the web scenario,” *Multimed. Tools Appl.*, Vol. 46, p.p. 47–69, 2010.

[27] O. Van Laere, J. Quinn, S. Schockaert, and B. Dhoedt, “Spatially Aware Term Selection for Geotagging,” IEEE Transactions on Knowledge and Data Engineering, Vol. 26, No. 1, Jan. 2014.

[28] https://blog.twitter.com/marketing/en_us/a/2015/new-research-twitter-users-love-to-watch-discover-and-engage-with-video.html, Retrieved, September 19, 2017.

[29] A. Doulamis, N. Doulamis and S. Kollias, “A fuzzy video content representation for video summarization and content-based retrieval,” Signal Processing, Vol. 80 p.p. 1049-1067, 2000.

[30] R. T. Ng and J. Han, “CLARANS: A method for clustering objects for spatial data mining,” IEEE Trans. Knowledge & Data Engineering, Vol. 14, No. 5, p.p. 1003-1016, 2002.