Graph-based social media story linking

Goncalo Marcelino, David Semedo, Joao Magalhaes

NOVA LINCS Faculdade de Ciências e Tecnologia Universidade NOVA Lisboa, Caparica, Portugal goncalo.bfm@gmail.com, df.semedo@campus.fct.unl.pt, jmag@fct.unl.pt

Abstract. This paper describes graph-based approaches to extract the visual elements from social media regarding the TRECVID Linking task story topics. We propose a sequential graph model and a fully connected graph.

1 Introduction

Today, with the large amounts of information available in the most varied contexts, being able to present compelling and informative narratives in an immediate and impactful fashion has revitalized the importance of the image. As a prime example of this, news media is focusing more and more on the usage of images to tell news stories, providing news in formats such as BBC's *In Picture*¹, where news are presented to the audience through selections of images.

Some authors have researched methods to automatically illustrate news with social media images [2]. However, the high amounts of information and images available, specially online, has also brought new challenges to the task of creating narratives through imagery. Rooted in the need to help solve these challenges and introduce new technologies into the millenia-old human tradition of creating stories with images, the TRECVID 2018 [1] Video linking task approaches visual storylines from a computation perspective, a problem yet to be solved by the research community.

In particular, the NOVA Search team focus on the problem of generating visual storylines to illustrate news pieces using social media content using graph methods to infer the visual story links.

2 Proposed framework

The proposed framework is composed of two modules: the first one aims to select all potential relevant content; the second module targets the creation of a sequence of video/images that provides a smooth visualization of the storyline:

 Retrieving relevant content. Ensuring the storyline is comprised of quality content isn't enough as illustrating a story with quality content that isn't relevant to the topic the story describes mutes the purpose of the illustration. Hence, the second module

¹ www.bbc.com/news/in_pictures

tackles the problem of retrieving relevant social media content to a particular story in an automated way. Taking as input the images filtered by the first model and a story to illustrate, this module finds candidate images to illustrate each segment of the story and provides them to the third and final module that composes the framework.

- Creating the storyline. As already described, a visual storyline is an ordered sequence of images and, as noted in Chapter 2, the way this images are ordered affects how they are perceived: from a human cognition point of view a visual storyline is expected to be semantically and visually coherent while providing an interesting narrative that unfolds over time. Taking the candidate images outputted by the second module as input, this final module is tasked with generating storylines with the candidate images that reflect these characteristics. As discussed in Chapter 2, although there is already research work on tasks such as semi-automated video editing, no works could be found on fully automated visual storyline generation. As such this task is also one that is one that is also yet to be tackled in literature.

In the next sections we will detail the implementation of the two modules.

3 Finding relevant story video/image material

In order to find relevant content to illustrate the individual segments of each story, we consider four different retrieval baselines. For each segment the rankings that result from the use of the baselines are then combined into a single one using Reciprocal Rank Fusion, parameterized with k = 60.

Text Retrieval The first and simplest of the methods proposed for segment's illustration, is a text retrieval baseline (BM25), that takes only into account the text associated with the social media documents. Given a set of social-media documents and a segment to be illustrated, we first apply stemming and stopword filtering to both the text of the tweets and the segment's text. Afterwards, BM25 is used to rank the posts containing images, according to their textual relevance to the segment's text. This approach is used as the basis for the remaining baselines.

Social Signals The previously described approach makes use of text retrieval only. This means that the quality of the visual content being used to represent a segment of a story is not taken into account. To tackle this problem, two baselines are proposed which use social signals as proxies of the users opinions, defining the quality and interestingness of the content. All posts in both baselines are originally ranked using BM25, as in the case of the Text Retrieval approach.

In the first baseline, referred to as *#Retweets*, the 20 first documents ranked by BM25 are re-ranked by the amount of times they where shared (e.g. "re-tweeted" in the case of Twitter). Alternatively, in the second baseline, referred to as *#Duplicates*, they are re-ranked by the number of times the image present in the post appears in the posts dataset.

Visual Semantics Visual content in social media documents may not match directly the corresponding textual content, which means that visual content can contain additional interesting information related to the event. Two additional baselines are proposed here which try to extend the information by exploiting the visual concepts of the images in each post.

- Concept Pool (CP): in this baseline, the VGG-16 convolutional neural network [5] is used to extract semantic concepts from the images present in the top 10 ranked posts, by *Text Retrieval*. These concepts are pooled together. Finally, the top 10 posts ranked by BM25 are re-ranked according to the number of concepts from this pool.
- Concept Query (CQ): this baseline is based on pseudo-relevance feedback with concepts extracted from images [3]. Semantic concepts are extracted from visual content in the top 5 ranked posts. These concepts are used to form a new query, which is then used to rank posts again. At this stage, the TF-IDF is used as the ranking algorithm.

Temporal Modelling Large scale events are usually composed of multiple sub-events that take place at different moments in time. Taking this insight into account a final baseline was proposed, referred to as *Temp. Modelling*, that prioritises content published closer (in time) to activity peak dates of documents related to the segment being illustrated.

Posts are first ranked using *Text Retrieval*. The number of posts published per day related to the segment being illustrated is then computed. At this point, the creation dates of the retrieved documents containing any of the words in the query are considered. A Kernel Density Estimator (KDE) is applied with a Gaussian kernel to the timestamps, resulting in a probability distribution over each date. The kernel bandwidth was fixed according to the method in [4]. The top 10 ranked documents by the Text Retrieval baseline are then re-ranked according to the KDE propability associated with their publication date.

4 Aiming for visual story coherence

Having tackled the problem of retrieving relevant content we now approach the task of optimizing for transition quality. To do so we deploy a machine learning approach supported by a novel formalization of the concept of transition.

4.1 Defining transition

To tackle the task of optimizing the transition quality between two pairs of images we first need a computationally valid approach to describe the concept of transition. From a non-computational, professional perspective, literature characterizes transitions based on the relations between semantic and visual characteristics of the images that compose them and by the ways in which these images interact. We emulate this approach proposing a novel formalization of transition based on the concept of distance. More

specifically we define the *transition distance* between two sequential images a and b of a visual storyline as:

$$(\forall c \in C, distance_c(feature_c(a), feature_c(b))) \tag{1}$$

where C is a set of image features under consideration, $feature_c$ is a function that returns feature c of an image, and $distance_c$ is a function that returns the distance between the same feature of two images. Hence, a transition between two images is formalized as a *transition distance*: a list of calculated distances between the features of said images.

4.2 Transition quality

Rating a transition between a pair of images according to its quality is a non-linear process that results from the interpretation of the features of the individual images and of the manner in which they interact. To tackle the automation of this process we resort to the regression version of Gradient Boosted Trees, defining the problem as one of predicting a rating given the *transition distance* of a pair. Aiming to build a robust model we propose the use of large set of features to compose the transitions distances between image pairs. These features are presented in Table 1. As training data we considered the transition quality ground truth provided in the context of this task.

We propose and evaluate two different main approaches each with two variants.

4.3 Sequencial

This fist approach considers transition quality individually, as opposed to attempting to ensure the creation of cohesive storylines as a whole. This means considering that the quality of transitions is impacted solely by the characteristics of the pair of images that compose it. Based on this approach we propose two variations.

Without relevance The first simpler alternative is optimized for creating storylines with the best transitions possible and is formally defined as follows: given a story with N segments

$$Story_N = (u_1, u_2, ... u_N)$$
 (2)

and list N sets of candidate images, each set G_i containing relevant, candidate images to illustrate story segment u_i :

$$Candidates_N = (G_1, G_2, \dots G_N) \tag{3}$$

We find $VisualStoryline_N = (j_1, j_2, ..., j_N)$ where $\forall i \in (1..N) :: j_i \in G_i$ that maximizes the function:

$$Quality = \sum_{i=2}^{N} t_{i-1,i} \tag{4}$$

where $t_{i,k}$ is the normalized quality of transition for the image pair j_{i-1} , j_i as predicted by the machine learning model.

We refer to this approach as Sequential (run 1).

c (feature name)	$distance_c(S1, S2)$	$feature_c(Si)$
Quality Difference	abs(f(S1) - f(S2))	A positive real value representing the aesthetic quality of the image.
Quality Sum	-abs(f(S1) + f(S2))	A positive real value representing the aesthetic quality of the image.
Environment	f(S1) = f(S2)	If the image represents a place outdoors or indoors .
Faces	abs(f(S1) - f(S2))	The number of faces in the image.
Scene Attributes	$\#(f(S1) \cap f(S2))$	The characteristics of a scene described in individual words.
Scene Category	$\#(f(S1) \cap f(S2))$	The most probable locations of a scene described in individual words.
Color Correlogram	jsd(f(S1) - f(S2))	A 16 bins 3D color correlogram in LAB color space.
Heat Map	$\sum abs(f(S1) - f(S2))$	A heat map of informative parts of the image.
Color Histogram	$\sum abs(f(S1) - f(S2))$	A 16 bins 3D histogram in LAB color space.
CNN Dense	$\sum abs(f(S1) - f(S2))$	A thing extracted from the last layer of a neural network.
Color Moment	euclidean(f(S1), f(S2))	Color moment in LAB color space.
Entropy	abs(f(S1) - f(S2))	A positve real value representing the entropy.
Concepts	$\#(f(S1) \cap f(S2))$	A set of image concepts extracted using VGG16.
#Edges	$\sum abs(f(S1) - f(S2))$	A vector of three positions with the number of horizontal, vertical and diag- onal edges, respectively.
Luminance	abs(f(S1) - f(S2))	A real value representing the luminance.
pHash	$\sum abs(f(S1) - f(S2))$	A Phash vector.

 Table 1. Transition features, and respective distance functions and descriptions.

Considering relevance The previous alternative considers only transition quality when generating visual storylines. Building on this first alternative we propose a second more complex one, this time considering the relevance scores of the pieces of content to the segments they are illustrating. Hence we directly optimizing for a function similar to the quality metric considered in this task.

Formalizing the method, for $Story_N$ and $Candidates_N$, as defined in the previous alternative, we find $VisualStoryline_N = (j_1, j_2, ..., j_N)$ where $\forall i \in (1..N) :: j_i \in G_i$ that maximizes the function:

$$Quality = 0.1 \cdot s_1 + \frac{0.9}{2(N-1)} \sum_{i=2}^{N} pairwiseQ(i)$$
(5)

$$pairwiseQ(i) = \underbrace{0.6 \cdot (s_i + s_{i-1})}_{\text{segments illustration}} + \underbrace{0.4 \cdot (s_{i-1} \cdot s_i + t_{i-1,i})}_{\text{transition}}$$
(6)

where s_i is the normalized relevance of j_i to u_i as calculated per the method described in Section 3.

We refer to this approach as Sequential with relevance (run 2).

4.4 Fully connected

The sequential approaches optimize for individual transition quality. However a visual storyline is consumed as a whole by it's viewers, not as a disconnected set of pairs. Consequently we posit a second approach designed to ensure cohesion between all elements of the generated visual storylines, leveraging the possibility that individual transition quality is affected by the remaining elements of the storyline they are part of. As with the sequential approach we establish two alternatives to this approach.

Without relevance First, again, optimizing only for transition quality, we simply consider the sum of the predicted quality of the transitions between all images in the story-lines, regardless if they appear in sequence or not.

Hence, for $Story_N$ and $Candidates_N$, we find $VisualStoryline_N = (j_1, j_2, ..., j_N)$ where $\forall i \in (1..N) :: j_i \in G_i$ that maximizes the function:

$$Quality = \sum_{i=2}^{N} \sum_{k \in \{1 < =k < =N \land k \neq i\}} t_{i,k}$$

$$\tag{7}$$

We refer to this approach as *Fully connected without relevance* (run 3).

Considering relevance This final alternative builds on the previous one, ensuring high transition quality between all pairs of the generated visual storylines, not just sequential pairs, while optimizing for the characteristics of the quality metric considered in this task.

Formalizing it, given $Story_N$ and $Candidates_N$, we find $VisualStoryline_N = (j_1, j_2, ..., j_N)$ where $\forall i \in (1..N) :: j_i \in G_i$ that maximizes the function:

$$Quality = 0.1 \cdot s_1 + \frac{(0.9)}{2(N-1)} \sum_{i=2}^{N} segmentQ(i)$$
(8)

$$segmentQ(i) = \underbrace{0.6 \cdot (s_i)}_{\text{segments illustration}} + \underbrace{0.4 \cdot \sum_{k \in \{1 < =k < =N \land k \neq i\}} (s_k \cdot s_i + t_{i,k})}_{\text{transitions}}$$
(9)

We refer to this approach as Fully connected with relevance (run 4).

5 Runs descriptions

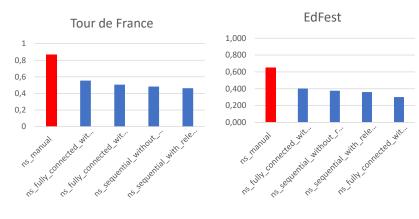
We submitted five runs for the video linking task:

- Run 1. This run creates a graph between documents of neighboring segments. The
 result is extracted by considering the best transitions between all documents of
 neighboring segments.
- Run 2. This run creates a graph between documents of neighboring segments. The
 result is extracted by considering the relevance of each document and the best transitions between all documents of neighboring segments.
- Run 3. This run considers a fully connected graph with all the documents retrieved by the baseline methods. The result is extracted by considering only the transitions between sequential segments.
- Run 4. This run considers a fully connected graph with all the documents retrieved by the baseline methods. The result is extracted by considering the relevance of each document and the transition between sequential segments.
- Run 5. This is a semi-automated run. The goal was to compare the automated runs with one where a search engine is used and the timeline organization of the story is done in a manual way.

6 Evaluation

Figure 1 illustrates the results of the described methods. We included a manual run for reference. Results show that the fully connected approach is the one that delivers the best story illustration. This is inline with the hypothesis that the global coherence of a story is essential. Also, while relevance plays in important role, with our methods the global coherence of the illustrations was more important than reranking the top 10 documents.

We believe that methods that deliver more relevance methods will be better, but considering the relevance of the top ranked documents is not as critical as finding the global visual coherence.





References

- G. Awad, A. Butt, K. Curtis, Y. Lee, J. Fiscus, A. Godil, D. Joy, A. Delgado, A. F. Smeaton, Y. Graham, W. Kraaij, G. Quénot, J. Magalhaes, D. Semedo, and S. Blasi. Trecvid 2018: Benchmarking video activity detection, video captioning and matching, video storytelling linking and video search. In *Proceedings of TRECVID 2018*. NIST, USA, 2018.
- D. Delgado, J. Magalhaes, and N. Correia. Automated illustration of news stories. In Semantic Computing (ICSC), 2010 IEEE Fourth International Conference on, pages 73–78. IEEE, 2010.
- J. a. Magalhães, F. Ciravegna, and S. Rüger. Exploring multimedia in a keyword space. In Proceedings of the 16th ACM International Conference on Multimedia, MM '08, pages 101– 110, New York, NY, USA, 2008. ACM.
- 4. D. W. Scott. On optimal and data-based histograms. Biometrika, 66(3):605-610, 1979.
- 5. K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.