#VisualHashtags: Visual Summarization of Social Media Events Using Mid-Level Visual Elements

Sonal Goel
IIIT-Delhi
sonal1426@iiitd.ac.in

Sarthak Ahuja
IBM Research, India
sarahuja@in.ibm.com

A V Subramanyam
IIIT-Delhi
subramanyam@iiitd.ac.in

Ponnurangam Kumaraguru
IIIT-Delhi
pk@iiitd.ac.in

ABSTRACT

The data generated on social media sites continues to grow at an increasing rate with more than 36% of tweets containing images making the dominance of multimedia content evidently visible. This massive user generated content has become a reflection of world events. In order to enhance the ability and effectiveness to consume this plethora of data, summarization of these events is needed. However, very few studies have exploited the images attached with social media events to summarize them using "mid-level visual elements". These are the entities which are both representative and discriminative to the target dataset besides being human-readable and hence more informative.

In this paper we propose a methodology for visual event summarization by extracting mid-level visual elements from images associated with social media events on Twitter (*#VisualHashtags*). The key research question is *Which elements can visually capture the essence of a viral event?*, hence explain its virality, and summarize it. Compared to the existing approaches of visual event summarization on social media data, we aim to discover #VisualHashtags, i.e., meaningful patches that can become the visual analog of a regular text hashtag that Twitter generates. Our algorithm incorporates a multi-stage filtering process and social popularity based ranking to discover mid-level visual elements, which overcomes the challenges faced by direct application of the existing methods.

We evaluate our approach on a recently collected social media event dataset, comprising of 20,084 images. We evaluate the quality of #VisualHashtags extracted by conducting a user-centered evaluation where users are asked to rate the relevance of the resultant patches w.r.t. the event and the quality of the patch in terms of how meaningful it is. We also do a quantitative evaluation on the results. We show a high search space reduction of 93% in images and 99% in patches after summarization. Further, we get a 83% of purity in the resultant patches with a data coverage of 18%.

MM'17, October 23–27, 2017, Mountain View, CA, USA.
 ISBN 978-1-4503-4906-2/17/10…$15.00
DOI: https://doi.org/10.1145/3123266.3123407

KEYWORDS

#VisualHashtags, Visual event summarization, Social media, Mid-level visual elements

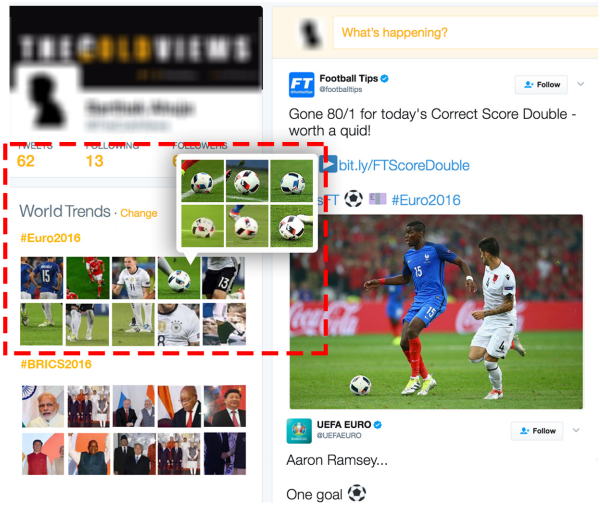


Figure 1: Highlighted in the image is a curated #VisualHashtag for the Euro2016 textual hashtag.

1 INTRODUCTION

Microblogging platforms, are widely used as a powerful medium to connect people during real-world events such as sports, politics, crisis situations, and so on. During these events users express their views on these platforms by posting textual and multimedia content. This data bolsters the opportunity for researchers and companies to analyze it, with the goal of understanding public opinion; tracking what's significant and most liked by users; understanding what's trending, etc. [4]. But the paramount size of this data raises challenges in its analysis and hence arises the need for data summarization, to obtain a compact description of the data.

Most of the existing summarization approaches focus more on the textual content as compared to a significant source of information: the multimedia content [1, 5, 19, 20]. Event summarization using visual archives has multiple applications; it reduces the size of

the corpus, increases the ability to better analyze and interpret the data, and guides the stakeholders to the interesting aspects of the data quickly. Images can convey much more information about a specific moment of an event as compared to text, which is typically short and is poorly written in multiple languages while images are naturally language independent and easier to understand [12]. Also, in the duration that a user spends to understand the gist of an image, one can read only one to four words [7].

However, new opportunities also bring along new challenges with itself. Some major challenges posed by exploiting social media images are (a) *uncertain quality*: images are posted without any quality guarantees, (b) *irrelevant content*: significant number of images (such as memes and screenshots) are irrelevant which adds noise to the content, and (c) *duplicity*: reposted and modified images using image processing techniques. Recent advances in vision make the domain highly conducive to bridge this gap and researchers have used visual content like images to summarize social media events [3, 4, 12, 18]. All of these studies summarize the social media events by either using images with text or images alone. To the best of our knowledge, mid-level visual elements have not been deeply explored for summarizing social media events.

The discovery of these mid-level visual elements is important as they cover the significant aspects of the images in the corpus of viral event and hence contribute to understand its virality. Here "virality" is synonymous with social media popularity in terms of the event appearing as trending topic. These mid-level visual elements are patches or structured regions of images that occur in several images with a certain degree of spatial consistency and can discriminatively summarize the event.

In this paper, we propose a methodology to discover meaningful mid-level visual elements (#VisualHashtags) from social media events that: (a) can capture its essence, (b) highlight its uniqueness, and (c) allow effective browsing of the event. These resultant #VisualHashtags are expected to more often correspond to full objects of the images. Figure 1, shows a diagrammatic representation of #VisualHashtags, where we summarize the images from Euro2016.

As mentioned above, one of the major properties of the mid-level visual elements is their discriminativeness. To account for this property in the summarization process, a relevant negative dataset is created, comprising of events against which this discriminativeness is required, as proposed by Doersch et al. [10]. In the context of social media, events falling in the same domain serve this purpose. Hence, the dataset is divided into two parts: (1) the positive set: containing images from the event which we want to summarize (2) the negative set: containing images from the remaining events belonging to same domain. For example, if we want to summarize EuroCup, then positive set will contain images from EuroCup and the negative set will contain images from other events related to sports domain like Wimbledon, Olympics, etc. We aim to discover meaningful visual elements, for which we handle duplicate image detection, detection of text in images, pruning object centralised patches using Deepmask [15], and then following the discriminative approach used by [10], select patches which occur frequently in the target event images and are also unique to that event. Finally, we rank the extracted patches using the social popularity score associated with their images and curate the results in form of #VisualHashtags.

As part of our evaluation and results, we summarize viral events on Twitter belonging to the sports and politics domain. Further, we show temporal analysis of visual summarization of US 2016 presidential elections before and after the declaration of results. Finally, we show a qualitative and quantitative evaluation of the summarization results.

The major contributions of this work are:

(1) We propose #VisualHashtags, a novel way to summarize images from social media events instead of the conventional method of only identifying key-images to represent the event.
(2) Our approach includes a multi-stage filtering process which when coupled with the basic methodology to discover mid-level visual elements leads to an improvement in coverage discussed in Section 5.3.

2 LITERATURE REVIEW

A substantial body of work exists in literature on the problem of textual summarization of social media events. Nichols et al. [14], summarize sports events on Twitter by using temporal cues like spikes in volume to determine key moments within an event. The authors first apply filtering techniques to tweets, such as removing spam, off-topic and non-English posts, before using a sentence ranking method to extract relevant sentences from the corpus of status updates describing each important moment within an event. In [6], the authors propose a probabilistic model for topic detection in Twitter, and use temporal correlation in the data to extract relevant tweets for summarization. Authors of [22] propose summarization of scheduled events on Twitter using a two-step approach, by first detecting sub-events through analysis of volume peaks and then selecting key tweets to describe each sub-event using a term frequency and Kullback-Leibler divergence weighting scheme. Chakrabarti et al. [5] propose to summarize event-tweets and give a solution based on learning the underlying hidden state representation of the event via Hidden Markov Models. In another study [17] the authors propose the task of personalized time-aware tweets summarization, selecting personalized meaningful tweets from a collection of tweets. They use user's history and collaborative social influences from social circles to infer dynamic probabilistic distributions over interests and topics. Authors in [6], propose a search and summarization framework to extract relevant representative tweets from a time-ordered sample of tweets to generate a coherent and concise summary of an event.

In recent years, the amount of visual data has increased tremendously and computer vision is one of the research fields that benefited from this. Researchers have also focussed on considering mutimedia content from social media to target problems like event summarization. In [18], Schinas et al. use both tweets and images to summarize an event. They reveal topics from a set of tweets as highly connected messages in a graph, whose nodes encode messages and whose edges encode their similarities. Finally, the images that best represent the topic are selected based on their relevance and diversity. Authors of [4] propose a social media imagery analytics system that processes and organize the images in more manageable way by removing duplicate, near-duplicate images and clustering images having similar content. In [12], the

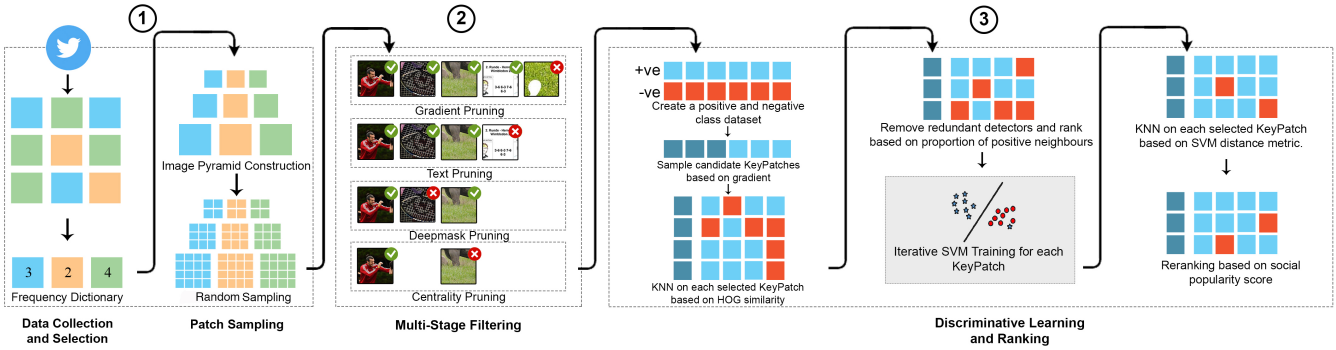


Figure 2: Overall flow of the approach. (1) First collect images from Twitter and remove duplicate images, use these images for patch sampling; (2) Apply multi-stage filtering to prune noisy and non-informative patches; (3) Apply discriminative learning to discover mid-level visual elements and rank them using social popularity score to finally present #VisualHashtag.

authors propose image selection and ranking method, which orders the most relevant images related to the event. To filter irrelevant images they eliminate memes, screenshot images and reaction images from the database and then detect near duplicates among them to increase diversity. Finally, they apply a ranking method to select a set of images that describes the event. Authors of [2] propose a generative probabilistic model-multimodal-LDA (MMLDA), to discover subtopics from microblogs by exploring the correlations among different media types. Based on the information achieved from MMLDA they design a multimedia summarizer to separately identify representative textual and visual samples and then form a comprehensive visualized summary.

All these social media event summarization use full images to describe events but none of them summarize an event by using mid-level visual elements. Discovery of these mid-level visual elements has been shown to be useful in various vision tasks like, image classification, including scene categorization and fine-grained categorization [11], action recognition [21]. Though there are studies that use mid-level image patches for data mining purposes, they are evaluated only on data that possess some definite patterns and do not consider the case of wild image dataset from social media. For example, Doersch et al. [10], collect data from Google Street View of different cities, and aim to automatically find the visual patches like windows, balconies, and street signs, that are most distinctive for a certain geo-spatial area. Another work by Rematas et al. [16], where the authors propose data-mining approach for exploring image collections by interesting patterns that use discriminative patches and further show the results on Pascal VOC and Microsoft COCO datasets. Towards the end, we aim to cover this research gap of summarizing viral social media events using mid-level visual elements or as we refer to throughout the paper, find meaningful image patches (in form of #VisualHashtags) that can capture the essence of a viral event on social media.

3 METHODOLOGY

3.1 Overview

We divide the overall approach in to three major steps (a) selecting unique images and sampling patches from these unique images (b) multi-stage filtering to extract meaningful patches from the sampled patches, and (c) discovering and ranking mid-level visual elements using discriminative learning incorporating social popularity score. Figure 2 shows the overall flow of the approach.

3.2 Data Collection and Image Selection

As mentioned earlier, in order to carry out the discriminative learning process, we first need to create positive and negative dataset. For our experiments, we collected data from June, 2016 to November, 2016, related to 7 events, 3 belonging to sports and 3 to politics category and 1 for doing temporal analysis on US 2016 presidential elections. To create this dataset, we collected data from Twitter using Twitter's search API [1], filtering tweets related to the keywords that were popular during the event. Further details of data collection is given in Table 1.

One of the major challenges posed by social media images is the large number of duplicate content posted. Though an image posted multiple times also adds to it's social importance, processing same images adds on to the computational complexity. Hence, we select only unique images for further processing. To extract unique images, we compute Perceptual hash [13] of the image using it's Difference Hash (dHash [2]) implementation. If the hash values of two images are same they are considered duplicates and grouped in a cluster. From each cluster only one unique image is selected and

[1] https://dev.twitter.com/rest/public/search
[2] https://pypi.python.org/pypi/ImageHash

Table 1: Details of Data Collection.

Event	Total Images	Unique Images	Category
EuroCup	3,489	827	Sports
Wimbeldon	3,229	1,327	Sports
Olympics	2,264	1,968	Sports
BREXIT	3,728	1212	Politics
BRICS	4,618	1,102	Politics
UNGA	2,756	1,572	Politics
US-Elections	98,813	5,000	Before-Election
US-Elections	218,289	5,000	After-Election

we also maintain a count of number of duplicate images present for each selected unique image. This count is later used as a score to rank the final results based on their social popularity. For the US Elections event, since the total images collected is much more than the other cases, we randomly sample top 5000 most retweeted unique images, to maintain the consistency with other events. The unique images column in Table 1 shows the reduction in number of images after pruning the duplicate images. On an average we are able to reduce 55% of duplicate images for all the events.

3.3 Sampling and Pruning Candidate Patches

After finding the unique images, we use them to sample random patches. We scale these images at various levels to randomly sample high-contrast patches of various resolution. The quality of patches plays a significant role in the process of generating a meaningful summary of an event. To prune the non-informative patches, we apply a multi-stage filtering process explained below. Stage-2 in figure 2 shows the sample patches pruned at each filtering stage.

As image gradients are used to extract information from images, we calculate the gradient magnitude of each patch by applying Sobel filter and take the mean value of its output. Only those patches that have a gradient value above a fixed threshold value (20 in our case, obtained experimentally) are allowed to pass to the second stage. With this, we are able to remove non-informative patches like just a plain background patch. We also discard very small patches, i.e. patches that have height or width less than 40px.

Now from the patches selected by the above filtering, we prune patches primarily containing text by using Pytesseract [3] (an OCR tool for python). If Pytesseract is able to detect text in a patch, we calculate the area covered by the textual region in the patch. If the text area is more than 50% of the patch area (i.e. a major portion in the patch is occupied by the text), we reject that patch. The reason to discard patches with a lot of text is that they mostly belong to memes or screenshot categories and are often less relevant to form the elements that would be useful to summarize the event. By the end of this step, we select patches where either Pytesseract is not able to detect text or the area of the textual region detected is less than 50% (chosen experimentally) of the patch area.

In the next level, we pass the resultant pruned patches from the previous step to a trained model, Deepmask [15]. The model is applied to an image and it generates a set of object masks, each with a corresponding objectness score. The objectness score defines the likelihood of the patch being centered on a full object. By experimentation, we set a threshold of 0.99, and if the objectness score of a patch is above the threshold, then the patch is passed to the next level.

Finally, we use the location of the masked object detected by Deepmask to identify if the object detected is in the centre of the patch or not. We calculate the centre of the bounding box of the masked object, and check if this point lies in the central window of the patch. The central window of the patch (having height h and width w) is defined as the area covered by a rectangular region ($h/2 \times w/2$) whose centre is same as the centre of the patch. If the centre of the detected object lies in the central window region, we assume that the masked object is located in the centre of the patch. By the end of this step, we select the patches where masked objects are located in the centre. These patches are known as candidate patches which will further be used in the discriminative learning approach to discover mid-level visual elements. This multi-stage filtering process can efficiently (a) discard noisy and less informative patches, (b) select patches which are more meaningful, often containing an object in the centre, and (c) reduce the number of unsuitable patches to large extent, aiding the linear SVM detector (discussed next) to learn from a better sample set and also speed up the overall process. Table 2 shows the reduction in the number of patches for different events when pruning is applied. On an average we are able to reduce 49.8% of patches after applying the multi-stage filtering process.

[3] https://pypi.python.org/pypi/pytesseract

3.4 Mining and Ranking #VisualHashtags

In the previous step, we sampled and pruned patches using various filters to discard noisy patches. Now, we find the patches from the target event, which occur frequently in the target dataset and are also discriminative to the target event. For this, we follow the *discriminative clustering* approach used in [10].

Each candidate patch from the target event (positive dataset) is taken as a seed patch, and its k nearest-neighbors are computed using HOG features [8], to form a cluster of similar patches. These seed patches are also known as detectors. Next, we rank the detectors based on the proportion of positive patches in their respective nearest neighbors sets and pick the top n detectors.

Although the aforementioned process forms a good method for sampling initial detectors, using HOG feature similarity to compare patches alone does not suffice to be a metric capable of creating visually coherent clusters [10]. Hence, this step is followed by the discriminative learning algorithm of iteratively training SVMs for each of the selected n detectors. The HOG representation of the patches is used as feature vectors in this SVM training process.

As referred in Algorithm 1, for each of the detectors, an SVM is trained with top k (5, in our case) positive nearest neighbors taken from previous KNN sampling and all the patches from the negative dataset taken as negative samples. This is followed with testing of the SVM learners on the positive set. After each round of this testing, the top k detections are added to the positive set of each cluster (K), i.e. initially the positive set size is k, after the next iteration it becomes $2k$, and so forth. This process is carried out for l iterations. After each iteration the detector is expected to improve it's capability to discriminate between the positive and the negative set. It should be noted, both the positive (P) and negative

Table 2: Number of patches selected at each filtering stage, and percentage of noisy patches pruned at the end.

Event	Initial Patches	Grad. Pruned	Text Pruned	Deepmask Pruned	Centr. Pruned	Reduce %
Euro	20,635	19,428	19,237	14,607	11,917	42%
Wimb	33,151	31,222	30,604	21,983	17,140	48%
Olymp	49,176	46,312	43,669	32,727	26,211	47%
BREXIT	30,264	28,828	27,677	17,544	13,937	54%
BRICS	27,526	26,363	25,124	18,431	14,821	46%
UNGA	39,276	38,108	36,229	24,021	19,849	49%

(N) sets are divided into l parts, and used in pairs. The training part of the algorithm takes place on one of these pairs and is tested on a different one. The number of iterations is decided based on experimentation, being 3 in our case.

The final set of detectors are then ranked based on the proportion of positive patches in the set of nearest neighbors, the nearest neighbors now are calculated based on the SVM score instead of the earlier HOG feature similarity, along with the social popularity score integrated according to the following equation:

$$score_i = \sum_{j=1}^{n} (-1)^c S_j \frac{n-j+1}{n} \quad (1)$$

Where score_i is the score of the detector, n is the number of nearest neighbors, c is the class of the nearest neighbor (1 for negative, 0 for positive), S_j is the frequency (number of duplicates) of the image to which the patch belongs. Finally, after ranking the detectors and their clusters based on the social popularity score of their corresponding images, we select top N clusters to summarize the target event, these resultant patches are known as #VisualHashtags.

Algorithm 1 Discriminative Learning

```
D1, D2, D3 ... Dn                          ▷ Set of n Detectors
K1, K2, K3 ... Kn                          ▷ Clusters of nearest neighbors
P = P1, P2, P3... Pn                       ▷ Positive dataset divided into l parts
N = N1, N2, N3 ... Nl                      ▷ Negative dataset divided into l parts
for i = 1 to n do
    Ki = HogBasedKNN(Di, P1, k)
end for
for i = 1 to l do
    P* = ChooseWithoutReplacement(P)
    N* = ChooseWithoutReplacement(N)
    for j = 1 to n do
        SVMj = trainSVM(Dj, Kj, N*)
        Kj = [Kj, topSVMDetections(SVMj, P*, k)]
    end for
end for
for j = 1 to n do
    Scorej = score(Dj,Kj)
end for
```

4 RESULTS AND ANALYSIS

In this section, we present the analysis and visual summary obtained, when our approach is applied on the dataset we collected.

4.1 Summarizing sports and political events

Figure 3 and 4 show the summarization results obtained on Sports and Politics events respectively. Due to lack of space, we present here 10 randomly selected mid-level visual elements from top 20 in the results.

As can be noted from the summary shown, the top visual elements that cover the essence of EuroCup contains football, player's jerseys, players expressing different kind of emotions, logo of the tournament, etc. Visual elements like tennis racket, logo of the tournament, patches of stadium, tennis court, and players in different playing positions covers the essence of Wimbledon. For Olympics the summary contains the patches of medals, logo of the event, scoreboards and images of some players. Analyzing the political events, in BREXIT, patches portraying flags of European Union, street view of Britain and people protesting are the influential visual elements discovered. For BRICS, the logo of BRICS Summit, flags of different participating countries, and pictures of representatives of each country forms the crux of the dataset. While in UNGA apart from the logo of UNGA, it seems that the dataset is dominated by the images of a representative, whose speech was apparently one of the most talked about speech in UNGA-2016.

4.2 Temporal analysis during Election-2016

A useful application of this approach, suitable for social media dataset is to evaluate the change in summaries during the course of the event. Here, we show the shift in the key-patches posted by the users at two distinct time instants of the recently conceived US 2016 Elections. We collect images posted in context of the US 2016 presidential elections, and divide the collection into two parts: (a) images posted before the election day (8th Nov), and (b) images posted after 8th Nov. Figure 5 shows the visual summary and the shift in dominating patches of the US Election2016 before and after the election day. As can be noted from Figure 5(a), most of the visual elements from prior-elections dataset contains patches of banners which are generally part of campaigns before the elections. While if we see Figure 5(b), the visual corpus after the elections seems to be mostly containing maps of the US showing the results of the elections in different parts of the country, along with patches depicting the results in graphs and figures. It can be noted that the textual patches in the prior-election dataset are the ones, which were either not identified under OCR pruning or the ones where the textual area is less than 50% of the patch area.

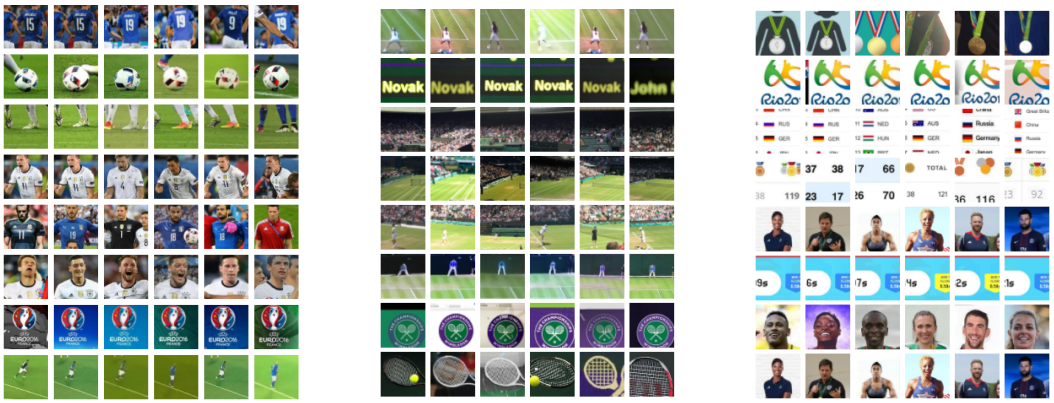
4.3 Linking #VisualHashtags

Another application of #VisualHashtags is to identify correlation among the detectors. The detectors from these visual elements can be linked based on the similarity of the images they belong to. Each detector is represented by a set of patches that are visually similar to it (a.k.a cluster of nearest neighbors), and each patch in the cluster belongs to an image from the event.

We visualize this linking by an undirected graph $G(V, E)$, where each node V is represented by the detectors in #VisualHashtag and E represents an edge between two detectors, if they are linked. We will use the terms node and detector interchangeably here.

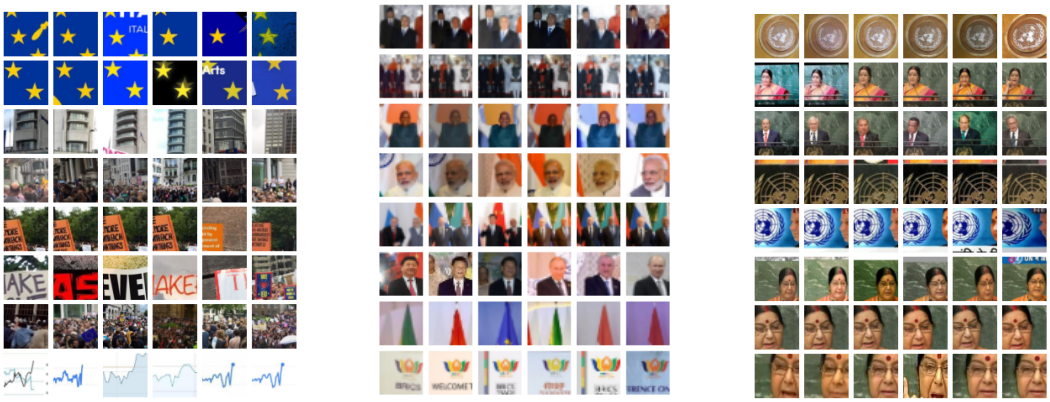
Table 3: Percentage of purity, coverage and search space reduction of images and patches offered after summarizing events.

Event	Purity	Coverage	SSR (Images)	SSR (Patches)
EuroCup	68%	25%	94%	98%
Wimbeldon	84%	11%	93%	99%
Olympics	90%	11%	90%	99%
BREXIT	68%	19%	94%	99%
BRICS	86%	27%	95%	98%
UNGA	100%	15%	92%	99%



(a) #VisualHashtags for EuroCup (b) #VisualHashtags for Wimbledon (c) #VisualHashtags for Olympics

Figure 3: Summarizing sports events for (a) EuroCup (b) Wimbledon (c) Olympics.



(a) #VisualHashtags for BREXIT (b) #VisualHashtags for BRICS (c) #VisualHashtags for UNGA

Figure 4: Summarizing politics events for (a) BREXIT (b) BRICS (c) UNGA.

To find if there exists a link between two nodes, we define a comparison function C. This function compares the corresponding images of the patches in the clusters of two detectors. For example, let there be two detectors D_x and D_y, comprising of n patches $\{P_{x1}, P_{x2}...P_{xn}\}$ and $\{P_{y1}, P_{y2}...P_{yn}\}$ in their cluster. Further, each patch in the cluster belongs to an image, forming image set of size m, $\{I_{x1}, I_{x2}...I_{xm}\}$ and $\{I_{y1}, I_{y2}...I_{ym}\}$. The comparison function returns the number of images that are similar (N), for the pair of detectors. In this implementation, we calculate the similarity of two images using perceptual hashing [13]. The equation below shows the implementation:

$$N = C(\,Similarity\,(I_{xi}, I_{yj})\,)\,, \quad \forall\,(i,j) \tag{2}$$

(a) Summarizing US Election2016 before the election day (b) Summarizing US Election2016 after the election day

Figure 5: Analysing the visual elements dominating the dataset before and after elections.

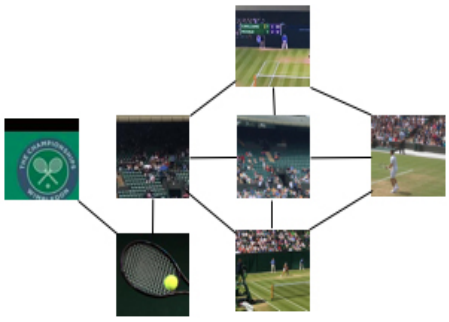


Figure 6: Graph showing connections between different patches signifying their co-occurrence in images.

Using the output of the comparison function, we calculate the proportion of patches in the cluster that belong to similar images (N/n), if this proportion is greater than a threshold t, then we can say that there is a link between the two detectors. This link between the two detectors signifies that, in most of the images these patches (objects in the patches) co-occur. This information can be further used to identify what kind of objects co-occur in the images of such viral events.

Figure 6 shows the pattern formed after linking the detectors from #VisualHashtags of Wimbledon. The links between patches of stadium, players and tennis court signifies that these patches (objects) mostly co-occur in the images of Wimbledon.

5 EVALUATION TECHNIQUE

In this section, we present an evaluation of the discovered #VisualHashtags for different datasets. We evaluate the approach at both quantitative and qualitative levels.

5.1 Quantitative Analysis

For quantitative analysis, we evaluate the summarization approach on three metrics (a) the discriminative quality of the patches in the result, (b) proportion of dataset covered by the resultant patches, and (c) the reduction in the search space that a user otherwise has to go through, thus saving on the time to analyse the data. Hence, we use a metric *purity* also used by [9], to evaluate the discriminative quality of the resultant mid-level visual elements. It is defined as the percentage of patches from the target event (positive dataset) in the result. We also calculate the *coverage*, which is defined as the percentage of the unique images covered in the dataset by the resultant patches. Finally, we calculate *search space reduction* (SSR) offered after the summarization process. Table 3 shows the purity, coverage and the percentage of search space reduction (SSR) for both images and patches, offered after summarizing each event. We observe a significant search space reduction of 93% in images and 99% in patches. We also observe a high purity of 83% on average and a mean coverage of 18% unique images.

5.2 Qualitative Evaluation

We follow the similar approach as [12, 18] to create relevance judgements for the #VisualHahtags selected to summarize different events through a user-centric evaluation. The group of annotators comprised 21 persons 20-30 years old. For each event, we ask 4 set of questions:

(1) The #VisualHashtag comprising of the top 10 patches from the final ranked result are shown to the users. The aim of this question is to check if the users are able to identify the event by just looking at the #VisualHashtag.

Task Description: Given the set of patches below, choose the most appropriate event which it summarizes.

Summary: EuroCup, Wimbledon, Olympics and UNGA are correctly identified by 100% of the users, while BRICS is accurately identified by 91% users and BREXIT by 95% of the people.

(2) Top 10 patches from #VisualHashtag are shown. The aim of this question is to find how many patches are *relevant* w.r.t. the event selected above.

Task Description: Select all the patches that can be distinctly linked with the event chosen above.

Summary: 51% of the users said atleast 4 out of 10 patches alone can be distinctly linked with the event. The Mean and standard deviation of the user's answers is shown in relevance section of table 4.

(3) The top 10 patches are shown to the users. The aim is to find how many patches correspond to full objects or meaningful parts of an image.

Task Description: Select all the patches that are *meaningful*, i.e. covering a meaningful part of an image.

Summary: For all the events on an average, more than half of the users said that atleast 50% of patches are meaningful. The Mean and standard deviation of the user's answers is shown in meaninfulness section of table 4.

(4) The #VisualHashtag along with the cluster of each patch (detector) is shown. The aim is to check the quality of clustering done.

Task Description: How many of the below rows demonstrate strong correlation (containing similar elements like faces/buildings etc) among their elements?

Summary: For all the events on an average, half of the users said 8 or more out of 10 mid-level visual elements show strong correlation in their clusters (nearest-neighbors).

Referring to the evaluation metric followed on user evaluations in [12], we also use the following same metrics for a more thorough qualitative assessment on the user evaluation for our results:

(1) **Precision (Pr@N)**: The percentage of patches among the top N that are relevant/meaningful to the corresponding event, averaged among all events. We calculate precision for N equal to 1, 5, and 10.

(2) **Success (S@N@D)**: The percentage of responses, where there exist at least D relevant/meaningful patches amongst the top N. We calculate success for N equal to 10 and D equal to 1, 3 and 5.

(3) **Mean Reciprocal Rank (MRR)**: Computed as 1/r, where r is the rank of the first relevant/meaningful patch returned, averaged over all events.

The value of all the three metrics mentioned above vary from 0-1, where higher values means better results. These metrics are computed for the two quality measures we aim to test in question 2 and 3 above- "relevance" and "meaningfulness" of the patch. As table 4 reads:

(a) For Pr@1, the precision on both quality-measures remains high for both politics and sports events when evaluating the top patch (i.e. for N=1), which is confirmed by evaluating their respective values of MRR as well. Further, if we see Pr@5, i.e. for N=5, close to half of the patches are considered relevant and meaningful in both types of events.

(b) Evaluating success, at N=1 (S@10@1), we observe that all the events have at least 1 relevant and meaningful patch thus the success rate is high for all the events in both sport and politics category. While, for N=3 (S@10@3), the success rate on both the measures for the events is close to 60% for sports and 46% for politics, i.e. at least 3 high-quality patches are generated close to half the time for all the events.

(c) In general, it can be observed that sports events reflect better results (in terms of relevance and meaningfulness) compared to the politics events. One of the reasons that can be attributed to this observation is that, one can easily connect patches like jerseys, balls, rackets, players, logos, etc to a sports event.

(d) Looking at the intersection (Mea.+ Rel.), we observe that there

Table 4: Precision, Success, MRR, Mean and Std. Dev., based on the qualitative analysis of the summarized events.

Quality-measures	Pr@1	Pr@5	Pr@10	S@10@1	S@10@3	S@10@5	MRR	Mean	S.Dev	Category
Relevance (Rel.)	0.94	0.57	0.45	1.0	0.75	0.48	0.95	4.5	2.4	Sports
Meaningfulness (Mea.)	0.95	0.58	0.44	1.0	0.79	0.43	0.97	4.3	2.0	Sports
Intersection (Rel.+ Mea.)	0.90	0.46	0.32	0.95	0.60	0.27	0.50	3.3	1.9	Sports
Relevance (Rel.)	0.79	0.54	0.41	1.0	0.57	0.37	0.87	4.1	2.6	Politics
Meaningfulness (Mea.)	0.79	0.58	0.41	1.0	0.71	0.38	0.87	4.1	2.1	Politics
Intersection (Rel.+ Mea.)	0.75	0.44	0.29	0.97	0.46	0.17	0.52	2.9	1.9	Politics

is a considerable overlap in meaningfulness of a patch and it's relevance indicating a correlation between the two.

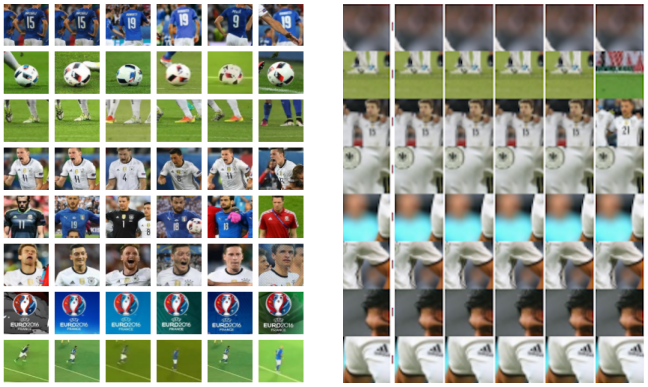
5.3 Comparison with basic discriminative method

In this section, we do a comparison of our approach with the basic discriminative method used by Doersch et al. [10]. We show the analysis on only the "EuroCup" dataset due to lack of space. Figure 7 shows the top-10 mid-level visual elements of EuroCup, obtained by our algorithm, filtering and discriminative (FILT_DISC) approach and the discriminative (DISC) methodology followed by [10].

We also do a quantitative comparison showing the difference in the purity-coverage (defined in section 5.1) values of the two approaches, when top 20 descriptors are chosen and the number of nearest-neighbors are varied from 5-25 with a step size of 5. Table 5 shows the results of quantitative analysis, where $P@n$ and $C@n$ are the purity and coverage values with n nearest-neighbors selected. As can be noted, when we use our approach (FILT_DISC), with a slight reduction in purity of patches there is a high jump in the dataset coverage as compared to the DISC approach, for all the nearest-neighbor values. As can be seen from Figure 7 (b), the reason of high purity in direct application of discriminative (DISC) approach is the presence of the patches from duplicate images in the social media data, unlike Figure 7 (a), where the patches in the nearest-neighbors are from different images hence, covering a wider view of images present in the dataset. However, in Figure 7 (b), the redundancy in results is quite high and the user experience is negatively impacted. Further, it reduces the coverage of unique images to which the top n nearest patches belongs, lowering the diversity of data in the summarization results.

6 CONCLUSION AND FUTURE WORK

In this work we present a methodology to visually summarize social media events using #VisualHashtags. To extract these #VisualHashtags, we start by sampling large number of random patches, and filter them based on their gradient value, the textual region proportion, and probability of it containing an object in the centre. Next, we use a discriminative approach to discover a set of patches which are both representative and discriminative to the event and rank them using social popularity score to summarize an event.



(a) Summarizing EuroCup using (FILT_DISC)

(b) Summarizing EuroCup using DISC

Figure 7: Comparing summary obtained by our approach FILT_DISC in (a) with DISC approach in (b).

We further show the application of this approach in analytics. After finding a #VisualHashtag for an event, patterns can be mined among patches by linking them based on their co-occurrence in similar images. We also show that performing summarization of an event at different time instances, generates #VisualHashtags representative of the temporal change that takes place during the course of an event.

We evaluate our results using both qualitative and quantitative methods on sports and politics datasets. At the end, we also show a comparison of our approach with basic discriminative method on social media data.

Currently, our approach is centered around events that contain images with relative stylistic coherence and uniqueness, and thus #VisualHashtags generated also focus on concrete entities. As future work, the technique can be modified to summarize more abstract phenomenon like violence, summer, etc. The mid-level patches obtained as a summary of a particular viral event, can be further generalised to pave way for finding higher-level image features that can cover the essence of an event. While the current approach needs to be re-run to generate #VisualHashtags at different time instances, dynamic re-summarization would be an interesting direction to explore, making it a more real-time system.

Table 5: Percentage of purity and coverage of the results with different nearest-neighbors for DISC and FILT_DISC.

Approach	P@5	C@5	P@10	C@10	P@15	C@15	P@20	C@20	P@25	C@25
DISC	100.0	0.77	100.0	1.11	94.3	1.37	85.5	2.03	79.0	2.57
FILT_DISC	93.0	10.17	88.0	20.80	77.0	22.72	68.0	24.41	67.0	25.36

REFERENCES

[1] Omar Alonso and Kyle Shiells. 2013. Timelines as summaries of popular scheduled events. In *Proceedings of the 22nd International Conference on World Wide Web*. ACM, 1037–1044.

[2] Jingwen Bian, Yang Yang, and Tat-Seng Chua. 2013. Multimedia summarization for trending topics in microblogs. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*. ACM, 1807–1812.

[3] Jingwen Bian, Yang Yang, Hanwang Zhang, and Tat-Seng Chua. 2015. Multimedia summarization for social events in microblog stream. *IEEE Transactions on Multimedia* 17, 2 (2015), 216–228.

[4] Paulo Cavalin, Flavio Figueiredo, Maıra de Bayser, and Claudio Pinhanez. 2016. Organizing Images from Social Media to Monitor Real-World Events. (october 2016). http://gibis.unifesp.br/sibgrapi16

[5] Deepayan Chakrabarti and Kunal Punera. 2011. Event Summarization Using Tweets. *ICWSM* 11 (2011), 66–73.

[6] Freddy Chong Tat Chua and Sitaram Asur. 2013. Automatic Summarization of Events from Social Media.. In *ICWSM*.

[7] Veronika Coltheart. 1999. *Fleeting memories: Cognition of brief visual stimuli*. Mit Press.

[8] Navneet Dalal and Bill Triggs. 2005. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, Vol. 1. IEEE, 886–893.

[9] Carl Doersch, Abhinav Gupta, and Alexei A Efros. 2013. Mid-level visual element discovery as discriminative mode seeking. In *Advances in neural information processing systems*. 494–502.

[10] Carl Doersch, Saurabh Singh, Abhinav Gupta, Josef Sivic, and Alexei Efros. 2012. What makes paris look like paris? *ACM Transactions on Graphics* 31, 4 (2012).

[11] Yao Li, Lingqiao Liu, Chunhua Shen, and Anton van den Hengel. 2016. Mining mid-level visual patterns with deep CNN activations. *International Journal of Computer Vision* (2016), 1–21.

[12] Philip J McParlane, Andrew James McMinn, and Joemon M Jose. 2014. Picture the scene...;: Visually Summarising Social Media Events. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*. ACM, 1459–1468.

[13] Vishal Monga and Brian L Evans. 2006. Perceptual image hashing via feature points: performance evaluation and tradeoffs. *IEEE Transactions on Image Processing* 15, 11 (2006), 3452–3465.

[14] Jeffrey Nichols, Jalal Mahmud, and Clemens Drews. 2012. Summarizing sporting events using twitter. In *Proceedings of the 2012 ACM international conference on Intelligent User Interfaces*. ACM, 189–198.

[15] Pedro O Pinheiro, Ronan Collobert, and Piotr Dollar. 2015. Learning to segment object candidates. In *Advances in Neural Information Processing Systems*. 1990–1998.

[16] Konstantinos Rematas, Basura Fernando, Frank Dellaert, and Tinne Tuytelaars. 2015. Dataset fingerprints: Exploring image collections through data mining. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4867–4875.

[17] Zhaochun Ren, Shangsong Liang, Edgar Meij, and Maarten de Rijke. 2013. Personalized time-aware tweets summarization. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. ACM, 513–522.

[18] Manos Schinas, Symeon Papadopoulos, Yiannis Kompatsiaris, and Pericles A Mitkas. 2015. Visual event summarization on social media using topic modelling and graph-based ranking algorithms. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*. ACM, 203–210.

[19] Beaux P Sharifi, David I Inouye, and Jugal K Kalita. 2013. Summarization of twitter microblogs. *Comput. J.* (2013), bxt109.

[20] Chao Shen, Fei Liu, Fuliang Weng, and Tao Li. 2013. A Participant-based Approach for Event Summarization Using Twitter Streams.. In *HLT-NAACL*. 1152–1162.

[21] LiMin Wang, Yu Qiao, and Xiaoou Tang. 2013. Motionlets: Mid-level 3d parts for human motion recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2674–2681.

[22] Arkaitz Zubiaga, Damiano Spina, Enrique Amigó, and Julio Gonzalo. 2012. Towards real-time summarization of scheduled events from twitter streams. In *Proceedings of the 23rd ACM conference on Hypertext and social media*. ACM, 319–320.