Pattern Recognition Letters
journal homepage: www.elsevier.com

Temporally Sorting Images from Real-World Events

Rafael Padilha**, Fernanda A. Andaló, Bahram Lavi, Luís A. M. Pereira, Anderson Rocha

Institute of Computing, University of Campinas, Campinas, SP, Brazil

ABSTRACT

As smartphones become ubiquitous in modern life, every major event — from musical concerts to terrorist attempts — is massively captured by multiple devices and instantly uploaded to the Internet. Once shared through social media, the chronological order between available media pieces cannot be reliably recovered, hindering the understanding and reconstruction of that event. In this work, we propose data-driven methods for temporally sorting images originated from heterogeneous sources and captured from distinct angles, viewpoints, and moments. We model the chronological sorting task as an ensemble of binary classifiers whose answers are combined hierarchically to estimate an image's temporal position within the duration of the event. We evaluate our method on images from the Notre–Dame Catedral fire and the Grenfell Tower fire events and discuss research challenges for analyzing data from real-world forensic events. Finally, we employ visualization techniques to understand what our models have learned, offering additional insights to the problem.

1. Introduction

Surveillance, satellite, and television media content have been a cornerstone for forensic investigators in the reconstruction and visual analysis of events, such as natural catastrophes and terrorist attacks. With the diffusion of smartphones and social media, the capturing and sharing of such events — which in the recent past could be made only by journalists and the mainstream media — can now be made by any citizen with unprecedented speed and reach. This diffusion grants the power of content generation virtually to all of us, spawning a vast flow of information that, combined with other multimedia sources, can potentially aid in the forensic investigation of such events. However, this information is shared in an unstructured way, so that temporal organization among the different sources is not known, making it directly unfeasible for sensitive purposes.

From a forensic point of view, it is essential to mine temporal evidence from available media to better understand the narrative of the event (i.e., the chronological order of what has happened) and the possible relations between the facts that compose it. Additionally, mining temporal knowledge is essential for fact-checking, as fake news can be produced by claiming that an image was captured in a different moment in time, helping to convey a different version of the facts.

(a) Grenfell Tower fire (London, 2017)

(b) Notre-Dame Cathedral fire (Paris, 2019)

Fig. 1. Images from events collected from social media. We look for visual clues to infer chronological order (e.g., differences in illumination or in the appearance of a scene), ignoring uninformative elements (e.g., bystanders and occluding structures). We kindly challenge the reader to sort these shuffled images in time (answer provided in the bottom of the next page).

Manually reconstructing the timeline of an event through the data is a challenging task. Besides requiring a deep understanding of the nature of the event, the large volume of available data is prohibitive for expert analysis. The number of media pieces collected from contemporary events easily surpasses the human capacity to organize and sort data within a reasonable time. In the case of the "Boston Marathon Bombings" event, in which two bombs exploded near the marathon finish line in 2013, Twitter was flooded with more than 700,000 mentions of the attack in less than two hours, including images and videos

**Corresponding author:
e-mail: rafael.padilha@ic.unicamp.br (Rafael Padilha)

captured at the event [20]. Not only would this be a time-consuming task, but also human annotation might even introduce bias to posterior analyses.

Even though images and videos recorded by popular devices might have timestamps embedded into their metadata, it is a common practice of social networks to remove such information when new content is uploaded. Even when present, they might not be trustworthy to infer chronological information, as they can be easily altered by freely available metadata editors (e.g., Metadata++ or EXIF Date Changer Lite).

In this sense, a way to overcome these problems is to sort media pieces by mining their visual information automatically. Methods in the literature reason about time by detecting specific pieces of visual evidence and understanding how they change as time progresses. Among explored evidence are motion of dynamic elements [15, 24], alterations in appearance and/or visual style of objects of interest [23, 10, 16, 17], and subtle variations in illumination, shadows, and natural color [8, 29]. Despite their results in estimating time-of-capture or chronological order of images, these methods require these particular visual elements to distill temporal knowledge from data.

Ideally, when analyzing an event, we want to benefit from as many available clues as possible, however, each event is a universe on its own and, consequently, not all visual evidence might be present or relevant for the chronological sorting. Besides that, existing techniques are often evaluated under controlled scenarios, whereas data that originated from real-world events is often noisy, presenting extreme illumination conditions, motion blur, poor quality, and varying degrees of occlusion from objects and bystanders. Additionally, as each media piece captures a distinct viewpoint, techniques must account for the possibility that a set of pictures might not record the same spatial position or even share a moment in time, even though they still depict the same event.

With this in mind, the goal of this work is to temporally sort images originated from real-world events automatically. We extend upon our previous work [13], in which we explored occlusion techniques to improve the temporal sorting of the Notre-Dame Cathedral Fire [4]. In this work, we approach this task holistically, capturing and modeling how the global appearance of the scene changes as time goes by. We propose a new data-driven method that breaks the overall duration of an event into smaller successive intervals and models the temporal relationships between them hierarchically.

Considering that real events often present a small amount of high-quality data, we also explore data augmentation techniques — MixUp [26], CutMix [25] and occlusion-based augmentation [27] — to increase the available data and improve the robustness of our models. We evaluate the proposed method in the Notre-Dame Cathedral Fire and the Grenfell Tower Fire [2] events (Figure 1), outperforming our previous method [13]. Finally, we rely upon visualization techniques [19] to peek under the hood of our models and interpret what they have learned, offering additional insights on the problem.

For both events from Figure 1, the correct order is **IV-II-III-I**.

2. Proposed Method

Considering the nature of the event being analyzed, it might be necessary to precisely define the moment an image was captured. On other occasions, it might be convenient to define if an image was captured before or after a specific time or if it corresponds to a sub-event. We propose to break the overall duration of an event into smaller successive intervals of time and model the temporal relationship between them. These intervals can be modeled as specific quanta of time (e.g., one-hour or one-day periods) or sub-events within the overarching event. *Our goal is to infer in which interval an input image was captured.*

Instead of approaching the inference problem as a multi-class classification, i.e., predicting the exact moment in time in which an image was captured within the event of interest, we explore a similar strategy to Martin et al. [12], modeling the relations between intervals with a series of *"Before vs. After"* (BvA) questions. As pointed by the authors, it is often easier to answer the question, "Was this image taken before or after this date?" rather than predicting the date directly.

In this line of thought, we train multiple binary *Before vs. After* classifiers, each optimized to answer whether an input image was captured before or after a particular point in time. The idea is to compare two contiguous groups of images in a binary fashion to break the problem into smaller, ideally easier problems. By focusing on a specific cutoff point in time, each BvA classifier will learn the subtle differences between neighboring intervals. Once individual BvA models are trained, their knowledge over the event's temporal semantics is aggregated for a more precise inference.

Considering an event, we split its duration into a set of n consecutive time intervals $T = [t_1, \ldots, t_n]$, with the ultimate goal of inferring interval $t_i \in T$ in which an unlabeled image was captured. To construct BvA classifiers, we set as cutoff points intervals t_j, resulting in classifiers C_j for $j \in \{1, \ldots, n-1\}$. During inference, an input image I is processed by each C_j and the corresponding probability p_j that I was captured before the end of interval t_j is annotated. Intuitively, $(1 - p_j)$ refers to the probability that I was captured after interval t_j.

We investigate two distinct approaches — Non-hierarchical and Hierarchical — to train each C_j, combine their answers, and finally decide during which interval an image was captured.

2.1. Non-hierarchical aggregation

We train each model C_j with all available training data, organized in a *before* class (for images located in intervals $[t_1, \ldots, t_j]$) and an *after* class (for images located in intervals $[t_{j+1}, \ldots, t_n]$). The number of samples per-class will depend on the related cutoff interval. To deal with this inevitable imbalance, we feed, to each C_j, input batches with an equal number of samples for both classes, randomly repeating elements from the class with fewest samples, if needed.

In the pipeline depicted in Figure 2, all classifiers are considered to produce the final prediction for I. The probability $P(t_j|I)$ that I was captured in an interval $t_j \in T$ is given by:

$$P(t_j|I) = \prod_{i=1}^{j-1}(1 - p_i) \prod_{i=j}^{n-1}(p_i). \quad (1)$$

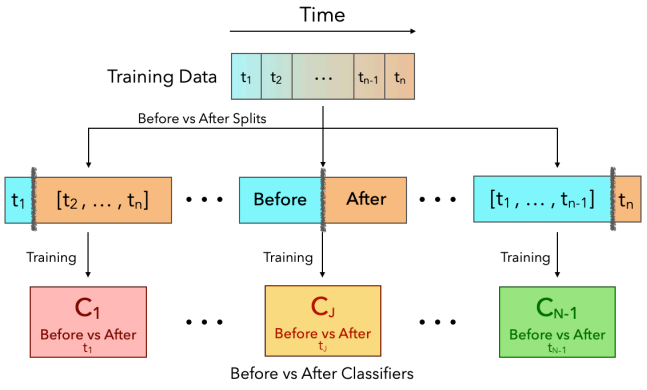

Fig. 2. Non-hierarchical pipeline. Each *Before vs. After* model is optimized with all available training data, assigned to *before* or *after* class depending on the corresponding interval. During inference, the answers of all classifiers are combined to output the final interval of capture.

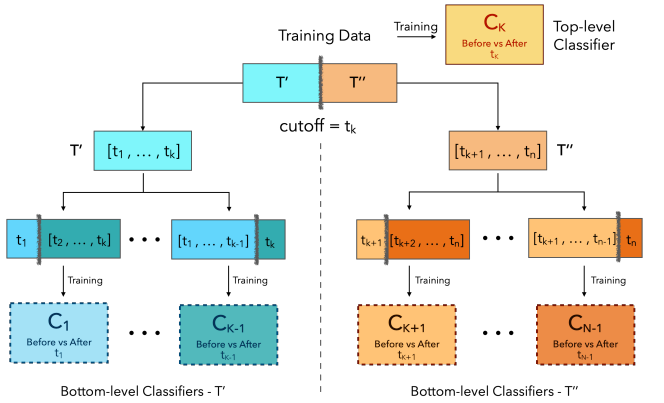

Fig. 3. Hierarchical pipeline. The top-level model is trained with all available data, whereas classifiers in the subsequent level are optimized with subsets T' or T'' of the data. During inference, the top-level model determines which group of classifiers will process the input image and estimate the interval it was captured.

In practice, this can be interpreted as the probability that I was captured *after* every interval prior to t_j, but *before* intervals following t_j. Finally, the predicted interval y is given by:

$$y = \underset{t_j \in T}{\text{argmax}}\ P(t_j|I). \tag{2}$$

2.2. Hierarchical aggregation

Instead of considering all intervals simultaneously, in this pipeline, we organize BvA classifiers in a hierarchical topology (Figure 3). We employ a top-level model to split intervals into two sets and train specialized BvA classifiers using only data from each set. By applying this *divide-and-conquer* approach, we want bottom-level classifiers to be more sensitive to slight visual changes in time, as they are trained with a smaller and closer set of intervals; thus, improving the overall inference.

Initially, a top-level BvA model C_k splits intervals T into sets $T' = [t_1, \ldots, t_k]$ and $T'' = [t_{k+1}, \ldots, t_n]$, for $1 < k < n-1$, assigning an input image I to be classified by bottom-level models from T' or T'', depending on the answer from C_k. We train C_k using all training data, whereas each bottom-level model C_j is trained using only data from the particular set (T' or T'') that interval t_j is located at.

Ideally, the top-level cutoff interval t_k should be selected considering the characteristics of the target event. As an example, when analyzing the Notre-Dame Cathedral Fire event (Section 4.1), we considered the fall of the cathedral's central spire — a drastic change in the building's appearance — as the cutoff sub-event; whereas on Grenfell Tower Fire event (Section 4.2), we selected the sunrise hour as the cutoff point, whereby average brightness of images shifts from darker to brighter values.

The probabilities $P(t_j|\{T', T''\}, I)$ that I was captured in an interval $t_j \in T$ are given by:

$$P(t_j|T', I) = \prod_{i=1}^{j-1}(1 - p_i) \prod_{i=j}^{k-1}(p_i), \tag{3}$$

$$P(t_j|T'', I) = \prod_{i=k+1}^{j-1}(1 - p_i) \prod_{i=j}^{n-1}(p_i). \tag{4}$$

Finally, the predicted interval y is given by:

$$y = \begin{cases} \underset{t_j \in T'}{\text{argmax}}\ P(t_j|T', I), & \text{if} \quad p_k > (1 - p_k), \\ \\ \underset{t_j \in T''}{\text{argmax}}\ P(t_j|T'', I), & \text{if} \quad p_k \leq (1 - p_k), \end{cases} \tag{5}$$

where p_k is the probability output by top-level classifier C_k, that I was captured until the interval t_k.

3. Datasets

We evaluated our methods in two real-world events: the Notre-Dame Cathedral Fire [4] and the Grenfell Tower Fire [2]. Both centered around a particular element of interest (cathedral and building, respectively), with available imagery collected from social media and traditional news sources. In this section, we describe each dataset and how they were organized.

3.1. Notre-Dame Cathedral Fire

In April 2019, a fire tore through the Notre-Dame Cathedral, an ancient Parisian architectural and religious symbol, devastating large parts of its structure and spire. People worldwide followed the tragic event through millions of images and videos shared by the media and everyday citizens.

Considering how the event developed in time, we can delineate it by a series of important episodes. After the fire started, it spread to the cathedral's spire. A few minutes later, the spire collapsed, and the fire continued burning until it was completed extinguished. By analyzing the event's description, four main episodes can be highlighted: *spire on fire*, *spire collapsing*, *fire continues on roof*, and *fire extinguished* (Figure 4).

The dataset[2] was collected from social networks, selecting public posts containing images captured only during and shortly after the event. These included memes, cartoons, compositions,

[2]The collected data is available in `https://doi.org/10.6084/m9.figshare.11787333`.

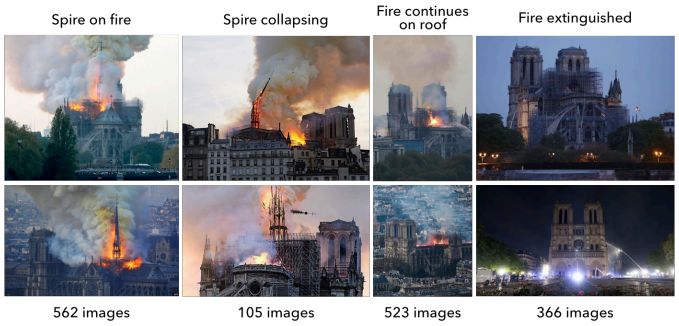

Fig. 4. Notre-Dame Cathedral during the fire that destroyed part of its structure in April 2019. Each column depicts one of the sub-events considered for the temporal ordering and number of images in each of them: *spire on fire*, *spire collapsing*, *fire continues on roof*, and *fire extinguished*.

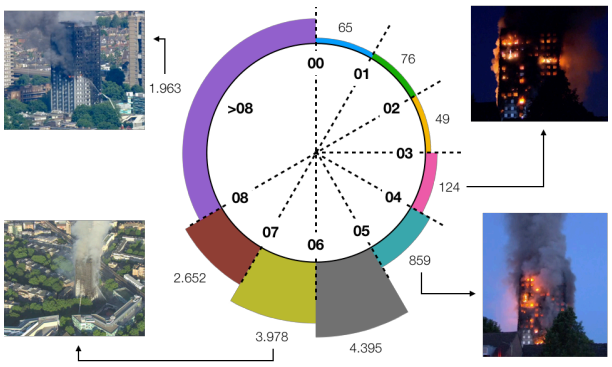

Fig. 5. The distribution of the available images in one-hour intervals. Note that, although the fire started before 01AM, most footage was captured after 04AM, probably when city inhabitants and media vehicles started broadcasting news about the event.

and images from the cathedral before the fire. To filter the *non-relevant* samples, we explored a semi-supervised approach based on the Optimum-path Forest theory [1], with features extracted with a CNN pre-trained on scene classification [28]. After mining the dataset for the set of relevant images, we also discarded duplicates and near-duplicates, which are abundant in this type of dataset due to the high rate of sharing content on social media. The cleaning process reduced available data using hashing methods [18, 22], which are effective in finding clusters of similar images. Refer to [13] for a complete description of the data collection and sanitization process.

3.2. Grenfell Tower Fire

The Grenfell Tower Fire [2] was an unprecedented fire that broke out in a residential building in London, in 2017, killing several people. The catastrophe was shared live by thousands of Londoners with their cameras and smartphones. The dataset contains roughly 150 videos of the event annotated concerning the cardinal direction of the visible facade and time of capture.

Each video had its keyframes extracted for a total of 14,155 frames. Even though it spanned for almost 24 hours, the first eight hours might be the most relevant for understanding what happened [3]. They comprise the start of the fire, the firefighters' efforts to control the flames and rescue entrapped people, as well as when news channels started broadcasting live footage. In this sense, we organized the available imagery in one-hour intervals, spanning from 00AM to 08AM. Images captured after 08AM were combined into a single *>08*AM interval. We present, in Figure 5, the distribution of images available for this event.

The footage is challenging to be analyzed, as it was captured from different angles, spanning different time frames, in varied resolutions. Differently from the previous event, half of the intervals are in the early morning, with poor illumination conditions that often lead to low-quality imagery. Additionally, the structural changes in the Grenfell Tower as time progresses are considerable subtler than of Notre-Dame Cathedral, which makes it harder to distinguish the order of intervals.

4. Experimental Analysis

In this section, we present empirical results of the temporal sorting of the Notre-Dame Cathedral Fire and Grenfell Tower Fire events. Considering each event's dataset, we randomly selected 50% of the images for training, taking care not to allow frames of the same video in different sets. We performed two-fold cross-validation using the data that was held out of the training set. We tracked the loss for each epoch, selected the best model based on one fold, and evaluated it on the other, repeating this process switching folds.

Each BvA classifier is an Inception-ResNet V2 [21] network, with pre-trained weights of the ImageNet dataset. We fixed all layers' hyperparameters, such as the number and size of filters, strides, and dropout factors. Due to the lack of data and with the goal of avoiding overfitting, we chose only to update the last convolutional and fully-connected layers, fixing the weights of previous layers. Each network was fine-tuned with batches of 32 images balanced per class with *Adam* optimizer [9] and a starting learning rate of 0.0001.

In [13], we evaluated occlusion-based data augmentation for sorting the Notre-Dame Cathedral Fire event in time, improving the accuracy and obtaining more discriminative features. We extend this evaluation for the methods proposed in this work, assessing the effectiveness of data augmentation techniques.

For all models, we applied the following augmentation during training: rotation in the range $[-10°, 10°]$, horizontal flip with 50% chance, and random scale and crop. For these last two techniques, the input image was resized so that its smaller dimension has the size in the range of $[340, 520]$ — keeping the aspect ratio — and random crop was applied for the final size of 299×299×3. In addition to that, we separately evaluate MixUp [26], CutMix [25] and random occlusion [27] techniques. For the latter, when including an image in the training batch, with a 50% chance, we randomly occluded a square region covering 10% to 45% of its size. For MixUp and CutMix, we employed the same formulation proposed by the authors.

Finally, at test time, the input image was resized so that its smaller dimension had a size of 340 pixels, and then a central crop was then applied for the final size of 299×299×3. Furthermore, we employ testing augmentation, creating ten versions of each testing image with rotation and horizontal flip, which are processed by the networks and their answers averaged.

We evaluated our methods in each dataset and reported the average balanced classification accuracy [5] for intervals $t \in T$:

$$Acc = \frac{1}{|T|} \sum_{t \in T} \sum_{I_j^t \in t} \frac{\mathbb{1}(y_j^t = t)}{|t|} \quad , \tag{6}$$

with $|T|$ representing the number of intervals in T, $|t|$ is the number of images in interval t, I_j^t the j-th image from t, and $\mathbb{1}(y_j^t = t)$ is equal to 1 if the prediction of I_j^t is correct, or 0 otherwise. Additionally, if our method incorrectly assigns an image to an interval, we wish that interval to be close to the ground-truth moment. To measure this, we also report the mean absolute error (MAE) and the *off-by-one* accuracy. The latter is calculated by considering predictions that miss by at most one interval of the ground truth as correct:

$$Acc_{off} = \frac{1}{|T|} \sum_{t \in T} \sum_{I_j^t \in t} \frac{\mathbb{1}(y_j^t \in \phi(t))}{|t|} \quad , \tag{7}$$

where $\phi(t)$ is the set of neighbor intervals of t including t, e.g., for the Grenfell Tower event, ϕ(01AM-02AM) = {00AM-01AM, 01AM-02AM, 02AM-03AM}. By averaging per-class classification accuracy, we obtain a measure of the performance that is not affected by the class imbalance in each dataset. The joint analysis of both metrics allow us to better evaluate if the methods are correctly capturing the ordinal relation between intervals.

4.1. Notre-Dame Cathedral Fire

To classify images temporally, we consider the images annotated with the sub-event in which they were photographed. Each sub-event has unique visual characteristics that are readily observable. For instance, for the first two episodes (*spire on fire* and *spire collapsing*), classifiers can learn to model the appearance of the spire. For the last two (*fire continues on roof* and *fire extinguished*), besides learning that the spire is no longer present, the model can focus on the fire.

In this sense, T = [*spire on fire*, *spire collapsing*, *fire continues on roof*, *fire extinguished*] and each interval $t_j \in T$ consists of one of the possible four sub-events. Thus, each BvA classifier C_j seeks to answer whether an input image was captured until that particular sub-event or after that.

Considering the Non-hierarchical aggregation, the answers of all three BvA classifiers are combined; and in the Hierarchical pipeline, we have a top-level classifier to first split the problem into "*until and after the spire is collapsing*". The idea is that the spire collapsing is a major sub-event that heavily affects the cathedral's appearance. It should be easier first to decide if an image was captured before/during this episode or after. Subsequently, specialized classifiers are used to decide between sub-events, which are harder to differentiate. In addition to both proposed aggregation methods, we also train a baseline CNN with images of the four significant sub-events in a multi-class setup, i.e., considering each sub-event as a class.

The results for the three different startegies are presented in Table 1, as well as the impact of applying MixUp, CutMix, and Random Occlusion as data augmentation techniques.

The Hierarchical approach outperforms both Non-hierarchical and multi-class models. As the overall task is being solved hierarchically, the method can benefit from solving an easier sub-task first — deciding if a captured sub-event deals with the cathedral's spire or not —, while the bottom-level classifiers focus on better discriminating the remaining classes. Additionally, applying MixUp, CutMix, or Random Occlusion increased the accuracy of the proposed methods, improving their robustness and decreasing the standard deviation.

Table 1. Accuracy Acc (%), off-by-one accuracy Acc_{off} (%) and mean absolute error MAE, followed by standard deviation, of the Notre-Dame Cathedral Fire event using three different strategies and data augmentation techniques. We also compare our approaches to [12] and [14].

	↑ Acc (%)	↑ Acc_{off} (%)	↓ MAE
Non-hierarchical	86.5 ± 2.7	**97.8 ± 0.1**	0.17 ± 0.04
Random Occlusion	90.5 ± 0.2	97.0 ± 0.2	0.16 ± 0.03
MixUp	91.0 ± 1.1	96.6 ± 0.4	0.15 ± 0.02
CutMix	90.5 ± 0.6	96.0 ± 0.5	0.18 ± 0.04
Hierarchical	90.1 ± 2.0	97.0 ± 0.4	0.14 ± 0.03
Random Occlusion	91.3 ± 0.1	96.7 ± 0.7	0.14 ± 0.03
MixUp	91.0 ± 0.9	96.4 ± 0.4	0.12 ± 0.02
CutMix	**92.1 ± 0.7**	96.5 ± 0.1	**0.11 ± 0.01**
Multi-class	88.3 ± 2.2	96.3 ± 0.1	0.15 ± 0.01
Random Occlusion	89.1 ± 1.2	96.7 ± 1.4	0.14 ± 0.01
MixUp	88.4 ± 2.2	94.9 ± 0.1	0.17 ± 0.03
CutMix	89.1 ± 1.5	96.6 ± 1.2	0.15 ± 0.01
Martin et al. [12]	55.9 ± 3.8	91.5 ± 0.7	0.57 ± 0.05
Palermo et al. [14]	73.8 ± 0.4	88.7 ± 0.7	0.40 ± 0.05

4.2. Grenfell Tower Fire

For this event, our goal is to estimate an one-hour window of time in which an image was probably captured. In this sense, we train a BvA classifier for each hour from 01AM to 08AM and evaluate both Non-hierarchical and Hierarchical pipelines to estimate the interval for unseen footage.

The models were fine-tuned with: images depicting the Grenfell Tower and background (*Full Image*), and manually cropped images encompassing only the building (*Building Crop*). Our intuition is that, by removing the background, the network will focus on identifying the burning patterns and correlating the evolution of the fire with intervals of time.

Additionally, we investigated the combination of building-crop and full-image models to assess their complementarity. In the *Score fusion* approach, we averaged the scores of CNNs trained for the same cutoff hour before aggregating all answers. We also explored (*Night-Day Fusion*) employing building-crop models for nighttime intervals (from 00AM to 03AM) — in which the background information seems less discriminative due to the lack of illumination — and full-image models for daytime (from 04AM onward). We report the results in Table 2.

Considering different input types, models optimized with building crops yielded better results than those trained with the full image. This indicates that the evolution of the fire and the appearance of the Grenfell Tower throughout the event play an essential role when estimating the time of capture.

As full-image networks also process background information, we expected their features would encode global characteristics of light and color and, thus, be complementary to building-crop features. Combining both models significantly improved the accuracy in the Hierarchical aggregation with no gain in the Non-hierarchical pipeline.

Table 2. Accuracy Acc (%), off-by-one accuracy Acc_{off} (%) and mean absolute error MAE, followed by standard deviation, for the Grenfell Tower Fire event. We evaluate the Non-hierarchical and Hierarchical pipelines, considering each binary CNN is trained with the full image or a crop depicting only the Grenfell Tower building, and also two different fusion approaches. We also compare them to [12] and [14] trained with full images.

	Non-hierarchical			Hierarchical		
	↑ Acc (%)	↑ Acc_{off} (%)	↓ MAE	↑ Acc (%)	↑ Acc_{off} (%)	↓ MAE
Full Image	41.8 ± 2.9	74.3 ± 1.4	0.94 ± 0.01	48.6 ± 7.6	78.2 ± 14.1	0.85 ± 0.23
Random Occlusion	39.5 ± 0.0	66.8 ± 2.5	1.04 ± 0.02	52.9 ± 11.1	78.3 ± 9.4	0.71 ± 0.31
MixUp	35.5 ± 6.5	65.8 ± 5.4	1.23 ± 0.31	43.1 ± 10.0	74.1 ± 16.7	0.90 ± 0.35
CutMix	37.6 ± 2.2	60.3 ± 2.8	1.21 ± 0.16	45.1 ± 2.4	79.2 ± 9.5	0.90 ± 0.23
Building Crop	46.4 ± 3.1	77.2 ± 4.7	0.84 ± 0.05	50.0 ± 2.4	71.0 ± 5.8	0.92 ± 0.02
Random Occlusion	**48.1 ± 1.7**	76.4 ± 0.1	**0.74 ± 0.09**	35.7 ± 3.0	72.8 ± 10.8	0.99 ± 0.11
MixUp	32.6 ± 4.2	65.8 ± 5.4	1.14 ± 0.12	43.6 ± 0.5	74.9 ± 10.4	0.88 ± 0.12
CutMix	33.7 ± 8.7	59.0 ± 4.1	1.32 ± 0.26	41.9 ± 0.8	73.8 ± 7.8	0.93 ± 0.11
Score Fusion	43.3 ± 3.8	75.2 ± 0.9	0.91 ± 0.00	**56.0 ± 7.5**	81.4 ± 11.2	0.77 ± 0.12
Random Occlusion	46.9 ± 2.0	73.1 ± 3.7	0.80 ± 0.07	52.6 ± 7.8	**83.8 ± 13.6**	**0.70 ± 0.26**
MixUp	34.1 ± 7.0	58.2 ± 7.2	1.20 ± 0.24	47.4 ± 0.6	78.3 ± 11.9	0.72 ± 0.29
CutMix	36.2 ± 6.6	61.3 ± 4.2	1.19 ± 0.26	47.4 ± 0.1	81.0 ± 11.5	0.78 ± 0.16
Night-Day Fusion	46.0 ± 3.5	77.5 ± 2.1	0.81 ± 0.01	52.3 ± 4.9	74.4 ± 5.7	0.83 ± 0.06
Random Occlusion	45.4 ± 1.3	**78.8 ± 0.8**	0.79 ± 0.06	43.3 ± 1.2	74.6 ± 10.8	0.83 ± 0.08
MixUp	40.8 ± 3.6	61.2 ± 4.9	1.05 ± 0.18	39.4 ± 2.2	72.9 ± 9.9	0.96 ± 0.14
CutMix	38.1 ± 1.6	61.7 ± 4.0	1.20 ± 0.17	44.3 ± 0.9	74.7 ± 9.1	0.87 ± 0.10
Multi-class						
Full Image	37.1 ± 2.4	75.0 ± 0.2	1.00 ± 0.02	-	-	-
Building Crop	37.9 ± 3.9	77.8 ± 4.2	0.99 ± 0.06	-	-	-
Score Fusion	41.1 ± 3.9	75.4 ± 0.9	0.94 ± 0.08	-	-	-
Martin et al. [12]	28.5 ± 9.7	70.1 ± 7.5	1.07 ± 0.49	-	-	-
Palermo et al. [14]	28.9 ± 2.9	67.5 ± 1.4	1.18 ± 0.21	-	-	-

Moreover, when comparing aggregation strategies, the Hierarchical approach achieved higher accuracy than the Non-hierarchical, indicating that training with images of similar brightness improved the final prediction. Despite achieving higher off-by-one accuracy in some cases, the standard deviation also increased, which might reflect the reduced number of samples when training each binary CNN in this pipeline.

4.3. Discussion

For both events considered in this work, the proposed methods outperformed the multi-class baseline, reinforcing the argument of Martin et al. [12] that modeling this task with an ensemble of binary classifiers allows each model to better capture subtle differences between consecutive intervals of time.

Our Hierarchical aggregation improved the sorting accuracy even further, indicating that such *divide-and-conquer* strategy is beneficial considering a cutoff point in time that split available data into groups with similar characteristics (e.g., brightness profile, or the presence/absence of the cathedral spire).

Despite its gain in classification accuracy, a drawback of this strategy is that, by splitting available data into bottom-level groups, each BvA model might only see a reduced amount of images, e.g., the nighttime models of the Grenfell Tower fire event (Figure 5). Consequently, such classifiers might be more prone to overfitting and require careful regularization strategies.

On the other hand, Non-hierarchical classifiers are optimized with the whole training data, having a global view of all intervals. However, models related to extreme intervals — i.e., near the beginning or end of the event — might be optimized with an imbalanced set due to the BvA formulation. In our experiments, we trained with balanced batches, oversampling underrepresented classes when needed.

Another strategy to deal with both regularization and class imbalance is data augmentation. The three techniques evaluated in this work — Random Occlusion, MixUp, and CutMix — offered better accuracy gains when applied to data from the Notre-Dame Cathedral event. We hypothesize this is due to

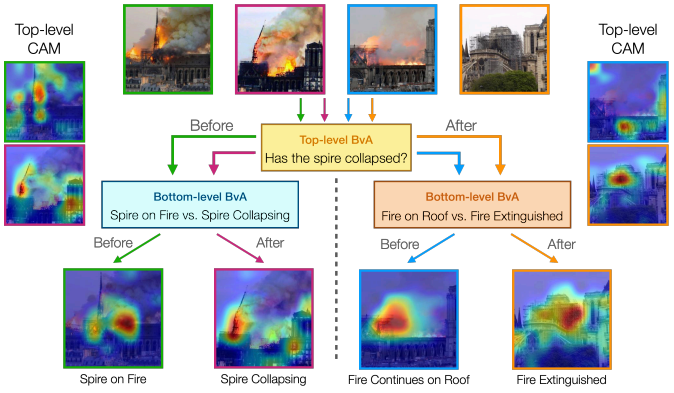

Fig. 6. Class activation maps [19] considering the best Hierarchical Aggregation setup for four input images (color-coded), one for each time interval, of the Notre-Dame Cathedral fire event. Warmer regions represent areas with more importance to the network decision. Each BvA classifier predicts whether an image was capture before or after a particular moment in time, and the final prediction is obtained by traversing the hierarchical model. The maps on the leaves are produced by the bottom-level classifiers, while the top-level classifier produces the side maps. Note that our classifiers give high importance to essential elements that determine the moment an image was captured, such as the spire collapsing or the fire in the roof.

sub-events of Notre-Dame Cathedral fire being naturally more distinguishable between each other (with drastic changes in the appearance of the cathedral) than the one-hour intervals of the Grenfell Tower fire. In this sense, the regularization effect offered by MixUp and CutMix techniques are undermined when combining images from intervals with similar characteristics (e.g., two nighttime intervals).

When compared to existing approaches [12, 14], our methods achieved superior performance in both forensic events. Their techniques employ handcrafted features that capture color and photo-generation artifacts that are not able to model the differences between intervals. More details of this evaluation can be found in the Supplementary Material.

To better understand what our models have learned, we generate the class activation maps [19] with the Hierarchical models for testing images of each event. The maps of the Notre-Dame Cathedral fire event (Figure 6) show that our models focus on specific elements of the scene that are informative to answer each "Before vs. After" question, such as the presence of the spire, fire spots on the roof and repair structures placed on top of the cathedral. We also note that the top-level model activates more broadly throughout the image — highlighting smoke regions and the facade of the cathedral — while bottom-level models present an activation peak on specific visual elements. Similar behavior can be seen in the class activation maps for the Grenfell Tower fire (Figure 7), in which the top-level BvA models use the appearance of the building to determine when the image was captured.

When considering the proposed framework for other events, the main modeling decision that might require an expert's intervention is the choice of number and duration of time intervals, as well as the top-level cutoff point for the Hierarchical method. These choices might be made based on characteristics of the event (e.g., the sub-episodes of Notre-Dame Cathedral Fire) and might lead to an improved temporal classification, as shown by our experiments. Nonetheless, an event-agnostic modeling can

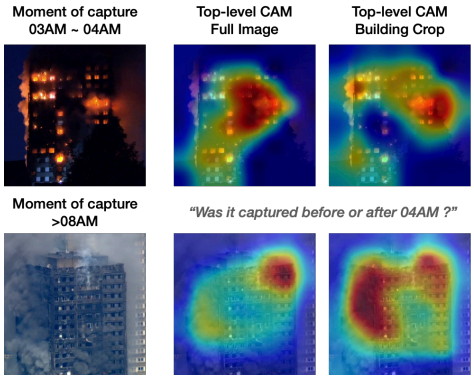

Fig. 7. Class activation maps [19] for the Grenfell Tower fire event, produced by the top-level classifiers of the *Score Fusion* Hierarchical Aggregation approach. Both classifiers use the appearance of the building and visible of fire spots through the windows to infer if each image was captured before or after 04AM.

be applied in most cases, e.g., by considering equally-spaced intervals throughout the event, similar to the one-hour division done in the Grenfell Tower Fire event.

Finally, in terms of model complexity, when compared to a single multi-class CNN, our approaches have a slightly higher training/inference processing time and increased model size. Instead of optimizing a model with N classes (one for each time interval), we train $N-1$ binary networks. At inference time, each image requires $N-1$ forward passes for the Non-hierarchical method and at most $\lceil \frac{N}{2} \rceil$ for the Hierarchical, against a single pass for the multi-class model. However, each of the networks takes only 60ms to process an image at inference time in a GeForce GTX 1080Ti GPU or 200ms on a single core of an Intel Xeon E5-2697 CPU with 2.30GHz. In terms of data requirement, we provide in the Supplementary Material a detailed exploration of the impact of decreasing the amount of training data available in the performance of our models.

4.4. Assessing generalizability: Dating Historical Pictures

The proposed approaches are not limited to forensic events. To demonstrate their generalizability, we also consider the task of estimating the decade in which a photograph was taken, following the problem presented by [14]. The dataset consists of 6,875 historical color pictures taken between 1930 and 1980, depicting varied types of scenes (Figure 8). Photographs from the same decade often present color patterns and artifacts innate of the photo-generation process of cameras from that age. In this sense, [12, 14] explored several handcrafted features that captured color information to predict the decade of each image.

We follow the same experimental protocol as [12, 14] to evaluate our approach. Each experiment was performed ten times, reporting the mean accuracy and mean absolute error (MAE) across all runs. In each round, images from each decade were randomly split into training and testing sets in the same ratio used by the authors. The training split was further split into 90% for training and 10% for validation.

Similar to our experiments for the forensics events, we used Inception-ResNetV2 as the base architecture, updating their weights of the last convolutional and fully-connected layers. Each model was optimized for 100 epochs, with batches of

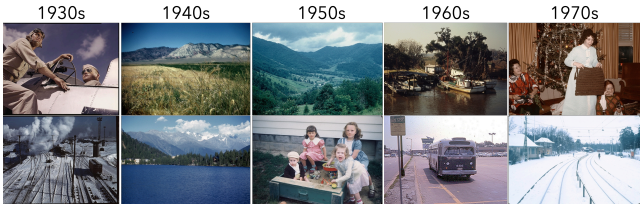

Fig. 8. Examples of images for each decade [14]. Even though the visual style of objects and fashion might help estimate the year of a picture, the most characteristic evidence lies in the overall color aspect of each decade. This is a challenging task, with untrained humans annotators achieving a classification accuracy of 26% [12].

Table 3. Accuracy and MAE, followed by standard deviation, in the task of dating historical color images. We compare our proposed approaches to existing methods, according to the results reproduced from [12].

Method	Accuracy (%)	MAE
Non-Hierarchical	**47.68 ± 2.37**	**0.81 ± 0.05**
Hierarchical	47.04 ± 1.52	0.89 ± 0.04
Multi-class	44.44 ± 3.02	0.93 ± 0.08
Palermo et al. [14]	44.92 ± 3.69	0.93 ± 0.08
Martin et al. [12]	42.76 ± 1.33	0.87 ± 0.05
Frank and Hall [7]	41.36 ± 1.89	0.99 ± 0.05
Cardoso and Costa [6]	41.32 ± 1.89	0.95 ± 0.04
Li and Lin [11]	35.92 ± 4.69	0.96 ± 0.06
Untrained human annotators	26.00	N/A

32 images, Adam optimizer, and initializing from weights pretrained in ImageNet. We tracked the loss for each epoch and selected the best checkpoint with respect to the validation set.

In the Non-hierarchical aggregation method, we trained four BvA models. Whereas, for the Hierarchical approach, a top-level classifier splits the decades into groups $\{1930s, 1940s, 1950s\}$ and $\{1960s, 1970s\}$ — i.e., *before or after the year 1960* — and bottom-level BvA models are trained within each group. Finally, we also optimize a multi-class network as a baseline for comparison. By doing so, we can better understand the differences in performance between our methods and previous techniques that employed handcrafted features. Table 3 presents the average accuracy and MAE along with standard deviation for all considered methods.

The multi-class CNN optimized for this problem achieves comparable performance to existing methods, indicating that the network could extract similar color information to the handcrafted features previously presented in the literature. As we can see, both proposed approaches outperformed prior techniques, with the Non-hierarchical aggregation peaking in accuracy, while considerably decreasing the MAE in this task. The results highlight that our models were able to better capture the appearance of photographs from each decade, while also learning the ordinal relation between classes.

Comparing both proposed methods, the Hierarchical approach was outperformed by the Non-hierarchical. We hypothesize that splitting the class space into two might not benefit bottom-level models as much in this task. The characteristics separating $\{1930s, 1940s, 1950s\}$ and $\{1960s, 1970s\}$ groups are considerably subtler than those of forensic events — e.g., the collapse of the spire in the Notre-Dame Cathedral or the illumination shift due to the sunrise in the Grenfell Tower

Fire. Considering this, we believe the Non-hierarchical models are able to capture these subtleties better in this particular problem, as each network has access to samples from all decades.

5. Conclusion

Once images of public events are captured and shared online, they lose most temporal ties to the real world. This makes it difficult to determine when each picture was taken, hindering posterior forensic analyses to reconstruct how the event unfolded.

In this work, we proposed a data-driven approach to chronologically sort images of real-world events. Our method breaks the event into small quanta of time, modeling their relation through "Before vs. After" classifiers, and hierarchically aggregates their answer to estimate the moment-of-capture of an image. We evaluated the proposed approach to sort images from two forensic events temporally — Notre-Dame Cathedral Fire [4] and the Grenfell Tower Fire [2] — significantly outperforming standard multi-class baselines.

In forensic scenarios, we often lack high-quality data to train robust deep networks fully. In this sense, data augmentation strategies are essential to increase variability and, consequently, model robustness virtually. We evaluated MixUp [26], CutMix [25], and random occlusion [27] during the training of our models and discussed — under the lens of forensic data — when these techniques work and fail.

As forensic methods make use of machine learning approaches, explaining a machine-made decision is essential for fairness and transparency. With this in mind, we inspected the visual elements considered important by our models, tying them to our knowledge of the scene.

The proposed techniques are flexible and can be easily adapted to new events. When doing so, a more accurate sorting is possible by grouping intervals with similar characteristics and training specialized models for each group. Even though the sorting accuracy might improve with more data, our experiments showed that augmentation techniques could be valuable when approaching forensic events with limited available data. Finally, temporal sorting methods can leverage forensic experts' knowledge, e.g., by labeling a subset of data or selecting intervals to be grouped, highlighting their importance in the forensic analysis pipeline.

Acknowledgment

We thank Forensic Architecture for the Grenfell Tower Fire data. This research was supported by São Paulo Research Foundation (FAPESP) [grant numbers 2017/12646-3, 2017/21957-2, 2018/05668-3 and 2018/16548-9] and the National Council for Scientific and Technological Development (CNPq).

References

[1] Amorim, W.P., Falcão, A.X., Papa, J.P., de Carvalho, M.H., 2016. Improving semi-supervised learning through optimum connectivity. Pattern Recognition 60, 72–85.

[2] BBC News, 2018a. Grenfell tower: What happened. https://bbc.in/2H4DhTc. Acc: 2019-01-29.

[3] BBC News, 2018b. The terrible speed with which the Grenfell fire spread. https://www.bbc.co.uk/news/uk-44381387. Acc: 2019-01-29.

[4] BBC News, 2019. Notre-dame: The story of the fire in graphics and images. https://bbc.in/2H1ZNvX. Acc: 2019-04-16.

[5] Brodersen, K.H., Ong, C.S., Stephan, K.E., Buhmann, J.M., 2010. The balanced accuracy and its posterior distribution, pp. 3121–3124.

[6] Cardoso, J.S., Costa, J.F., 2007. Learning to classify ordinal data: The data replication method. J. of Machine Learning Res. 8, 1393–1429.

[7] Frank, E., Hall, M., 2001. A simple approach to ordinal classification, in: Eur. Conf. on Machine Learning (ECML), pp. 145–156.

[8] Jacobs, N., Roman, N., Pless, R., 2007. Consistent temporal variations in many outdoor scenes, in: IEEE Conf. on Comput. Vision and Pattern Recognition (CVPR), pp. 1–6.

[9] Kingma, D., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

[10] Lee, S., Maisonneuve, N., Crandall, D., Efros, A., Sivic, J., 2015. Linking past to present: Discovering style in two centuries of architecture, in: IEEE Int. Conf. on Computational Photography (ICCP), pp. 1–10.

[11] Li, L., Lin, H.T., 2007. Ordinal regression by extended binary classification, in: Advances in Neural Inform. Process. Syst. (NIPS), pp. 865–872.

[12] Martin, P., Doucet, A., Jurie, F., 2014. Dating color images with ordinal classification, in: ACM Int. Conf. on Multimed. Retrieval, p. 447.

[13] Padilha, R., Andaló, F.A., Rocha, A., 2020. Improving the chronological sorting of images through occlusion: A study on the notre-dame cathedral fire, in: IEEE Int. Conf. on Acoustics, Speech, and Signal Process. (ICASSP).

[14] Palermo, F., Hays, J., Efros, A.A., 2012. Dating historical color images, in: Eur. Conf. on Comput. Vision (ECCV), pp. 499–512.

[15] Pickup, L.C., Pan, Z., Wei, D., Shih, Y., Zhang, C., Zisserman, A., Scholkopf, B., Freeman, W.T., 2014. Seeing the arrow of time, in: IEEE Conf. on Comput. Vision and Pattern Recognition (CVPR), pp. 2035–2042.

[16] Salem, T., Workman, S., Zhai, M., Jacobs, N., 2016. Analyzing human appearance as a cue for dating images, in: IEEE Winter Conf. on Appl. of Comput. Vision (WACV), pp. 1–8.

[17] Schindler, G., Dellaert, F., 2010. Probabilistic temporal inference on reconstructed 3D scenes, in: IEEE Conf. on Comput. Vision and Pattern Recognition (CVPR), pp. 1410–1417.

[18] Schneider, M., Chang, S.F., 1996. A robust content based digital signature for image authentication, in: IEEE Int. Conf. on Image Process. (ICIP), pp. 227–230.

[19] Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization, in: IEEE Int. Conf. on Comput. Vision (ICCV), pp. 618–626.

[20] Stern, J., 2013. Boston marathon bombing: The waves of social media reaction. ABC News: Good Morning America, https://abcn.ws/2ILGc3r. Acc:2019-06-02.

[21] Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A., 2017. Inception-v4, Inception-ResNet and the impact of residual connections on learning, in: AAAI Conf. on Artificial Intelligence, pp. 4278–4284.

[22] Venkatesan, R., Koon, S.M., Jakubowski, M.H., Moulin, P., 2000. Robust image hashing, in: IEEE Int. Conf. on Image Process. (ICIP), pp. 664–666.

[23] Vittayakorn, S., Berg, A.C., Berg, T.L., 2017. When was that made?, in: IEEE Winter Conf. on Appl. of Comput. Vision (WACV), pp. 715–724.

[24] Wei, D., Lim, J., Zisserman, A., Freeman, W.T., 2018. Learning and using the arrow of time, in: IEEE Conf. on Comput. Vision and Pattern Recognition (CVPR), pp. 8052–8060.

[25] Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y., 2019. Cutmix: Regularization strategy to train strong classifiers with localizable features, in: IEEE Int. Conf. on Comput. Vision (ICCV), pp. 6023–6032.

[26] Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D., 2017. mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412.

[27] Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y., 2017. Random erasing data augmentation. arXiv preprint arXiv:1708.04896.

[28] Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A., 2017a. Places: A 10 million image database for scene recognition. IEEE Trans. on Pattern Anal. Mach. Intell. .

[29] Zhou, H.Y., Gao, B.B., Wu, J., 2017b. Sunrise or sunset: Selective comparison learning for subtle attribute recognition. arXiv preprint arXiv:1707.06335.