VISIOCITY: A New Benchmarking Dataset and Evaluation Framework Towards Realistic Video Summarization

Vishal Kaushal
Department of Computer Science and Engineering
Indian Institute of Technology Bombay, India
vkaushal@cse.iitb.ac.in

Suraj Kothawade
Department of Computer Engineering
University of Texas at Dallas, USA
suraj.kothawade@utdallas.edu

Rishabh Iyer
Department of Computer Science
University of Texas at Dallas, USA
rishabh.ayer@utdallas.edu

Ganesh Ramakrishnan
Department of Computer Science and Engineering
Indian Institute of Technology Bombay, India
ganesh@cse.iitb.ac.in

Abstract

Automatic video summarization is still an unsolved problem due to several challenges. The currently available datasets either have very short videos or have a few long videos of only a particular type. We introduce a new benchmarking video dataset called VISIOCITY (VIdeo SummarIzatiOn based on Continuity, Intent and DiversiTY) which comprises of longer videos across six different categories with dense concept annotations capable of supporting different flavors of video summarization and other vision problems. Secondly, human reference summaries necessary for supervised video summarization techniques are difficult to obtain, especially for long videos. We explore strategies to automatically generate multiple reference summaries from indirect ground truth present in VISIOCITY. We also present a study of different desired characteristics of a good summary and argue that evaluating a summary against one or more human summaries and using a single measure has its shortcomings. We propose an evaluation framework for better quantitative assessment of summary quality which is closer to human judgment. We present insights into how a model can be enhanced to yield better summaries and demonstrate the effectiveness of our recipe in doing so as compared to some of the representative state of the art techniques tested on VISIOCITY. We release VISIOCITY and invite researchers to test their algorithms on VISIOCITY benchmark.

1 Introduction and Motivation

The unprecedented rise in the amount of video data has also made it difficult to consume them. This has given rise to the need for automatic video summarization techniques which aim at producing much shorter videos without significantly compromising on the key information contained in them. Consequently there has been a lot of work pushing the state-of-the-art for newer algorithms and model architectures [9, 31, 30, 32, 8] and datasets [6, 21, 24]. However the literature also talks of a few fundamental challenges that need to be addressed before we have a more realistic video summarization that works in practice. In this work we take steps towards addressing the following challenges:

1https://visiocity.github.io/

Dataset: Almost all recent techniques have reported their results on TVSum [24] and SumMe [6] which have emerged as benchmarking datasets of sorts. However, since the average video length in these datasets is of the order of only 1-5 minutes, they are far from being effective in real-world settings. While there have been several attempts at creating datasets for video summarization, they either a) have very short videos, or b) they have very few long videos and often of only a particular type. A large dataset with a lot of different types of full-length videos with rich annotations to be able to support different techniques was one of the recommendations in [27], is still not a reality and is clearly a need of the hour [9]. We introduce VISIOCITY to address this need. Furthermore, different flavors of video summarization, for example, query focused video summarization [28, 23], are often treated differently and on different datasets. With its rich annotations VISIOCITY can lend itself well to other flavors of video summarization and also other computer vision video analysis tasks like captioning or action recognition. Also, since the videos span across different well-defined domains, VISIOCITY is suitable for more in-depth domain specific studies on video summarization [27, 21].

Reference summaries for supervised learning: Supervised techniques tend to work better than unsupervised techniques [9] because of learning directly from human summaries. However, as video summarization is quite subjective there is no single ‘right’ answer and two human summaries could thus be quite different in their selections [10] [18]. Thus, more the number of human summaries, better is the learning. Unfortunately, for long videos different human summaries with diverse characteristics and of different lengths are difficult to obtain. In this work we explore strategies to automatically generate ground truth reference summaries which can be used to train a model. Further, most supervised learning approaches rely on the use of a combined ground-truth summary [30] [3] but in doing so they cannot fully explore the learning potential of such architectures [11] [32]. The necessity to deal with different kind of summaries in different ways was also observed by [27]. [11] [32] use this argument to advocate the use of unsupervised approaches. We leverage VISIOCITY to demonstrate that better results can be achieved when a supervised model learns from individual ground truth summaries using multiple loss functions each measuring deviation from different desired characteristics of summaries.

Evaluation: A video summary is typically evaluated by comparing it against all human summaries (for e.g., avg F1 or AFI) or nearest human summary (for e.g., max F1 or MF1). However, a good candidate may get a low score just because it was not fortunate to have a matching human summary, a likely scenario in case of long videos. We propose an evaluation framework where a summary is assessed on its own merit using the rich annotations in VISIOCITY. Also, a typical measure used is F1 [31] [8] which has its limitations. For example, it is not designed to measure aspects like continuity and diversity. Also, due to the segmentation used as a post processing step in typical video summarization pipeline, even random summaries can get good F1 scores [18]. Hence, instead of over dependence on one measure, we propose using a suite of measures to capture various aspects of a summary like continuity, diversity, redundancy, importance etc.

<table>
<thead>
<tr>
<th>Name</th>
<th># Videos</th>
<th>Avg Duration</th>
<th># Categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>SumMe</td>
<td>25</td>
<td>2m 39s</td>
<td>-</td>
</tr>
<tr>
<td>TVSum</td>
<td>50</td>
<td>4m 11s</td>
<td>10</td>
</tr>
<tr>
<td>MED</td>
<td>160</td>
<td>2.5m</td>
<td>15</td>
</tr>
<tr>
<td>UTE</td>
<td>4</td>
<td>254m 26s</td>
<td>1</td>
</tr>
<tr>
<td>Youtube 1</td>
<td>50</td>
<td>1m 39s</td>
<td>-</td>
</tr>
<tr>
<td>Youtube 2</td>
<td>50</td>
<td>2m 54s</td>
<td>-</td>
</tr>
<tr>
<td>Tour20</td>
<td>140</td>
<td>3m</td>
<td>-</td>
</tr>
<tr>
<td>TV Episodes</td>
<td>4</td>
<td>45m</td>
<td>1</td>
</tr>
<tr>
<td>LOL</td>
<td>218</td>
<td>30-50m</td>
<td>1</td>
</tr>
<tr>
<td>VISIOCITY</td>
<td>67</td>
<td>55m</td>
<td>6</td>
</tr>
</tbody>
</table>

Table 1: VISIOCITY dataset vs other existing datasets.

2 Related Work

Datasets: Currently available datasets for video summarization either have very short videos or have few long videos of only a particular type. Table 1 compares VISIOCITY with other existing datasets for video summarization. MED Summaries Dataset [21] consists of 160 annotated videos of length 1-5 minutes, with 15 event categories like birthday, wedding, feeding etc. The annotation comprises of segments and their importance scores. TVSum [25] consists of 50 videos (average length 4 minutes) from 10 categories with importance scores provided by 20 annotators for each 2 second snippet. The videos correspond to very short events like ‘changing vehicle tires’, ‘making sandwich’ etc. Though number of categories in TVSum and MedSummaries appear to be large, the notion of categories there is of events, like ‘making a sandwich’ or ‘attempting bike tricks’, quite
different from different domains in VISIOCITY with an intent of studying the characteristics of summaries of different types of videos like sports or TV Shows. The UT Egocentric Dataset [14] consists of long and annotated videos captured from head mounted cameras. However, though each video is very long, there are only 4 videos and they are of one type, i.e. egocentric. SumMe [7] consists of 25 videos with an average length of about 2 min. The annotation is in form of user summaries of length between 5% to 15%. Each video has 15 summaries. The VSUMM dataset [2] consists of two datasets. Youtube consists of 50 videos (1-10 min) and OVP consists of 50 videos of about 1-4 min from the Open Video Project. Each video has 5 user summaries in the form of set of key frames. Tour20 [19] consists of 140 videos with a total duration of 7 hours and is designed primarily for multi video summarization. It is a collection of videos of a tourist place. The average duration of each video is about 3 mins. TV Episodes dataset [29] consists of 4 TV show videos, each of about 45 mins. The total duration is 3 hours. A recent dataset, UGSum52 [15] offers 52 videos with 25 user generated summaries each. LOL [4] consists of online eSports videos from the League of Legends. It consists of 218 videos with each video being between 30-50 mins. The associated summary videos are 5 - 7 mins long. While this dataset is significantly larger compared to the other datasets, it is limited only to a single domain, i.e. eSports. [22] have extended the UTE dataset to 12 videos and have provided concept annotations, but they are limited to only egocentric videos and do not support any concept hierarchy. The scores annotations, as in TVSum etc. are richer annotations, but are limited only to importance scores. VISIOCITY on the other hand comes with dense concept annotations for each snippet. To the best of our knowledge, VISIOCITY is one of its kind large dataset with many long videos spanning across multiple categories and annotated with rich concept annotations for each snippet.

Techniques for Automatic Video Summarization: A lot of past work exists for automatic video summarization for example, using submodular functions [31,8,11,8,12], LSTMs [31], reinforcement learning [32] and attention models [9,8].

Evaluation: Typically every new technique of video summarization came with its own evaluation criteria [13,14,5] making it difficult to compare results across different techniques. More recently, computing overlap between groundtruth and generated summaries reported by F-measure has become the standard framework for video summary evaluation. Yet others prefer to evaluate a summary in the text domain as text is better at capturing higher level semantics [29,20]. This also forms the motivation behind our proposed evaluation measures. However, our measures are different in the sense that a summary is not converted to text domain before evaluating. Rather, how important its selections are, or how diverse its selections are, is computed from the rich textual annotations in VISIOCITY. This is similar in spirit to [22], but there it was done only for egocentric videos.

3 VISIOCITY Dataset

Videos: VISIOCITY is a diverse collection of 67 videos spanning across six different categories: TV shows (Friends), sports (Soccer), surveillance, education (Tech-Talks), birthday videos and wedding videos. The videos have an average duration of about 50 mins. Summary statistics for VISIOCITY are presented in Table 2. Publicly available Soccer, Friends, TechTalk, Birthday and Wedding videos were downloaded from internet. TV shows contains videos from a popular TV series Friends. They are typically more aesthetic in nature and professionally shot and edited. In sports category, VISIOCITY contains Soccer videos. These videos typically have well-defined events of interest like goals or penalty kicks and are very similar to each other in terms of the visual features. Under surveillance category, VISIOCITY covers diverse settings like indoor, outdoor, classroom, office and lobby. The videos were recorded using our own surveillance cameras. These videos are in general very long and are mostly from static continuously recording cameras. Under educational category, VISIOCITY has tech talk videos with static views or inset views or dynamic views. In personal videos category, VISIOCITY has birthdays and wedding videos. These videos are typically long and unedited. The videos are available to see and download from the project website at https://visiocity.github.io/

Annotations The ground truth in VISIOCITY is not direct in form of the user summaries, but indirect in form of concepts marked for each snippet. Being at a higher level, indirect ground truth can be seen as a ‘generator’ of ground truth summaries and thus allows for multiple solutions (reference summaries) of different lengths with different desired characteristics and is easy to scale. It also makes the annotation process more objective and easier than asking the users to directly produce reference ground truth summaries.
Concepts are a carefully selected list of verbs and nouns based on the type of the video and are given importance ratings based on the knowledge of the particular domain. **The concepts are organized in categories instead of a long flat list.** Example categories include ‘actor’, ‘entity’, ‘action’, ‘scene’, ‘number-of-people’, etc. Categories provide a natural structuring to make the annotation process easier and also support for at least one level hierarchy of concepts for concept-driven summarization.

In addition to concepts, we ask annotators to group those consecutive snippets as *mega-events* which together constitute a cohesive event. For example, a few snippets preceding a goal in a soccer video, the goal snippet and a few snippets after the goal snippet together would constitute a *mega-event*. A model trained to learn importance scores (only) would do well to pick up the *goal* snippet. However, such a summary will not be very pleasing to watch because what is required in a summary in this case is not just the ball entering the goal post, but the build up to this event and probably a few snippets as a followup. Thus, this notion of mega events helps us to model the notion of continuity.

**Textual Annotations vs Ratings or Scores as Indirect Ground truth:** While past work has made use of other forms of indirect ground truth like asking annotators to give a score or a rating to each shot [24], using textual concept annotations offers several advantages. First, especially for long videos, it is easier and more accurate for annotators to mark all keywords applicable to a shot/snippet than for them to tax their brain and give a rating (especially when it is quite subjective and requires going back and forth over the video for considering what is more important or less important). Second, when annotators are asked to provide ratings, they often suffer from chronological bias. One work addresses this by showing the snippets to the annotators in random order [25] but it doesn’t work for long videos because an annotator cannot remember all of these to be able to decide the relative importance of each. Third, the semantic content of a snippet is better captured through text [29, 20]. This is relevant from an ‘importance’ perspective as well as ‘diversity’ perspective. As noted earlier, two snippets may look visually different but could be semantically same and vice versa. Text captures the right level of semantics desired by video summarization. Also, when two snippets have the same rating, it is not clear if they are semantically same, or they are semantically different but equally important. Textual annotations bring out such similarities and dissimilarities more effectively. Fourth, as already noted, textual annotations make it easy to adapt VISIOCITY to a wide variety of problems.

**Annotation Protocol:** A group of 13 professional annotators were tasked to annotate videos (without listening to the audio) by marking all applicable keywords on a snippet/shot through a python GUI application developed by us for this task. It allows an annotator to go over the video unit by unit (shot/snippet) and select the applicable keywords using a simple and intuitive GUI. It provides convenience features like copying the annotation from a previous snippet, which comes in handy where there are a lot of consecutive identical snippets, for example in surveillance videos. Special caution was exercised to ensure high quality annotations. Specifically, the guidelines and protocols were made as objective as possible, the annotators were trained through sample annotation tasks, and the annotation round was followed by two verification rounds where both precision (how accurate the annotations were) and recall (whether all events of interest and continuity information has been captured in the annotations) were verified by another set of annotators. Whatever inconsistencies or inaccuracies were found and could be automatically detected, were included in our automatic sanity checks, which were run on all annotations.

### Table 2: Key Statistics of VISIOCITY.

<table>
<thead>
<tr>
<th>Domain</th>
<th># Videos</th>
<th>Duration (min, max, avg)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Soccer</td>
<td>12</td>
<td>(37, 122, 64) 12.8 hours</td>
</tr>
<tr>
<td>Friends</td>
<td>12</td>
<td>(22, 26, 24) 4.8 hours</td>
</tr>
<tr>
<td>Surveillance</td>
<td>12</td>
<td>(22, 63, 53) 10.6 hours</td>
</tr>
<tr>
<td>Educational</td>
<td>11</td>
<td>(15, 122, 67) 12.28 hours</td>
</tr>
<tr>
<td>Birthday</td>
<td>10</td>
<td>(20, 46, 30) 5 hours</td>
</tr>
<tr>
<td>Wedding</td>
<td>10</td>
<td>(40, 68, 55) 9.2 hours</td>
</tr>
<tr>
<td>All</td>
<td>67</td>
<td>(26, 75, 49) 54.68</td>
</tr>
</tbody>
</table>

Table 2: Key Statistics of VISIOCITY. Third Column is in minutes (min, max, avg).

### Proposed Evaluation Framework

Literature talks about certain desirable good characteristics of a video summary [8, 13, 14, 16, 30, 32]. For example, a good video summary is supposed to be diverse (non-redundant), continuous or visually pleasing (without abrupt shot transitions), representative of the original video and contain important
or interesting snippets from the video. In what follows, we dive deeper into these characteristics and propose measures to assess the candidate summaries on those characteristics.

**Diversity:** A summary which does good on diversity is non-redundant. It contains segments quite different from one another. different could mean different things in terms of content alone (i.e. one doesn’t want two similar looking snippets in a summary) or in terms of content and time (i.e. one doesn’t want visually similar consecutive snippets, but does want visually similar snippets that are separated in time) or in terms of the concepts covered (one does not want too many snippets covering the same concept and would rather want a few of all concepts). In surveillance videos for example, one would like to have a summary which doesn’t have too many visually similar consecutive and hence redundant snippets, but does have visually similar snippets that are separated in time. For instance, consider a video showing a person entering her office at three different times of the day. Though all three look similar (and will have identical concept annotations as well), all are desired in the summary. With regards to the quantitative formulation, we define the first flavor of diversity as: $Div_{sim}(X) = \min_{i,j \in X} d_{ij}$ where $X$ is a subset of snippets. $d_{ij}$ is IOU measure between snippets $i$ and $j$ based on their concept vectors. For the other two flavors of diversity, we define diversity clustered: $Div_{time}(X) = \sum_{i=1}^{\left| C \right|} \max_{j \in X \cap C} r_j$ where $C$ are the clusters, which can be defined over time ($Div_{time}$) (all consecutive similar snippets form a cluster) or concepts ($Div_{concept}$) (all snippets covering a concept belong to a cluster) and $r_j$ is the importance rating of a snippet $j$. When optimized, this function leads to the selection of the best snippet from each cluster. This can be easily extended to select a finite number of snippets from each cluster instead of the best one.

**MegaEventContinuity:** element of continuity makes a summary pleasurable to watch. Since only a small number of snippets are to be included in a summary, some discontinuity in the summary is expected. However, the less the discontinuity at a semantic level, the more pleasing is the summary to watch. There is a thin line between modelling redundancy and continuity when it comes to visual cues of frames. Some snippets might be redundant but are important to include in the summary from a continuity perspective. To model the continuity, VISIOCITY has the notion of mega-events as defined earlier. To ensure no redundancy within a mega event, the mega-event annotations are as tight as possible, meaning they contain bare minimum snippets just enough to indicate the event. A non-mega event snippet is continuous enough to exist in the summary on its own and a mega event snippet needs other adjacent snippets to be included in the summary for semantic continuity. We measure mega-event continuity as follows: $MegaCont(X) = \sum_{i=1}^{E} r_{\text{mega}}(M_i) \left| X \cap M_i \right|^2$ where, $E$ is the number of mega events in the video annotation, $r_{\text{mega}}(M_i)$ is the rating of the mega event $M_i$ and is equal to $\max_{s \in A} r(s). A$ is the annotation of video $V$, that is, a set of snippets such that each snippet $s$ has a set of keywords $K^s$ and information about mega event, $M$ is a set of all mega events such that each mega event $M_i$ $(i \in 1, 2, \cdots E)$ is a set of snippets that constitute the mega event $M_i$. 

**Importance / Interestingness -**

This is the most obvious characteristic of a good summary. For some domains like sports, there is a distinct importance of some snippets over other snippets (for e.g. score changing events).

This however is not applicable for some other domains like tech talks where there are few or no distinctly important events. With respect to the annotations available in VISIOCITY, the importance of a shot or snippet is defined by the ratings of the keywords of a snippet. These ratings come from a mapping function which maps keywords to ratings for a domain. The ratings are defined from 0 to 10 with 10 rated keyword being the most important and 0 indicated an undesirable snippet. We assign ratings to keywords based on their importance to the domain and average frequency of occurrence. Given the ratings of each keyword, rating of a snippet is defined as $r_s = 0$ if $\exists i : r_{K^s_i} = 0$, and $r_s = \max_i r_{K^s_i}$ otherwise. Here $K^s$ is the set of keywords of a snippet $s$ and $r_{K^s_i}$ is the rating of a particular keyword $K^s_i$. Thus, importance function can be defined as: $\text{Imp}(X) = \sum_{s \in X \cap A} r(s)$

Note that when both importance and mega-event-continuity is measured, we define the importance

<table>
<thead>
<tr>
<th>Measure</th>
<th>Expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>DiversitySim (DS)</td>
<td>$\min_{i,j \in X} d_{ij}$</td>
</tr>
<tr>
<td>Diversity(Time/Concept) (DT/DC)</td>
<td>$\sum_{i=1}^{\left</td>
</tr>
<tr>
<td>Mega Event Continuity (MC)</td>
<td>$\sum_{i=1}^{E} r_{\text{mega}}(M_i) \left</td>
</tr>
<tr>
<td>Importance (IMP)</td>
<td>$\sum_{s \in X \cap A \cap M} r(s)$</td>
</tr>
</tbody>
</table>
only on the snippets which are non mega-events since the mega-event-continuity term above already
takes care of the importance of the mega-event snippets.

As discussed earlier, since there are multiple "right" answers with varying characteristics, we hy-
ypothesize that these are orthogonal characteristics and vary across different human summaries. For
example, one human (good) summary could contain more important but less diverse segments while
another human (good) summary could contain more diverse and less important segments depending
on the intent behind summarization or user subjectivity. Also, in assessing summaries, one measure
could be more relevant than another depending on the type of the video. For example in sports videos
because of well-defined events of interest, importance is more relevant in evaluating a summary. We
verify our hypotheses experimentally. Motivated by this, we propose using a suite of measures as
defined above instead of overly depending on any one of them. The measures are computed using
the annotations present in VISIOCITY. We summarize them in Table 3. We propose that a true
and wholesome assessment of a candidate summary can only be done when this suite of measures
(including the existing measures like F score) are used. Results and observations from our extensive
experiments corroborate this fact.

5 Ground Truth Summaries for Supervised Learning

In practice, it is difficult to acquire many human summaries with diverse characteristics, especially
for long videos. We explore strategies to automatically generate the reference ground truth summaries
of desired lengths using the annotations present in VISIOCITY. Specifically, we use the above
proposed assessment measures as scoring functions and maximize them to get the desired ground
truth summaries. We note that maximizing a particular scoring function would yield a summary rich
in that particular characteristic, but it may fall-short on other characteristics. For example, a summary
maximizing importance will capture the goals in a soccer video, but some snippets preceding the goal
and following the goal will not be in the summary and the summary will not be visually pleasing
(example illustration at https://visiocity.github.io/). Hence, a weighted mixture of such measures
need to be maximized to arrive at optimal yet diverse reference summaries. This composite scoring
function (weighted mixture) takes an annotated video (keywords and mega-events defined over
snippets/shots) and generates a set of candidate (ground-truth) summaries which supervised semi
supervised summarization algorithms can use. Mathematically, given $V$, a set of snippets of a video
$V$, let $score(X)$ be defined as:

$$score(X, \lambda) = \lambda_1 MegaCont(X) + \lambda_2 Imp(X) + \lambda_3 Div_{sim}(X) + \lambda_4 Div_{time}(X) + \lambda_5 Div_{concept}(X)$$

This scoring function is parameterized on $\lambda$’s and is approximately optimized via a greedy algo-

This algorithm [17] to arrive at the ground truth summaries. Different configuration of $\lambda$s generates different
summaries. We use the notion of Pareto optimality to arrive at optimal configurations. Pareto optim-
ality is a situation that cannot be modified so as to make any one individual or preference criterion
better off without making at least one individual or preference criterion worse off. Beginning with a
random element (a possible configuration of the lambda) in the pareto-optimal set, we iterate over
remaining elements to decide whether a new element should be added or old should be removed,
or new element should be discarded. This is decided on the basis of the performance on various
measures. A configuration is better than another when it is better on all measures, otherwise it is not.

We verify experimentally that the automatic ground truth summaries so generated are at par with the
human summaries both qualitatively and quantitatively. We use them in training the models tested on
VISIOCITY.

6 Towards A New State of the Art

Following [3] we formulate the problem of automatic video summarization as a subset selection
problem where a weighted mixture of set functions is maximized to produce an optimal summary
for a desired budget. Specifically, given a video $V$ as a set of snippets $Y$, the problem reduces to
picking $y \subseteq Y$, which maximizes our objective such that $|y| \leq k$, $k$ being the budget. That is, $y^* =
\arg\max_{y \subseteq Y, |y| \leq k} o(x_v, y)$, where $y^*$ is the predicted summary, $x_v$ the feature representation of the
video snippets and $o(x_v, y)$ is the weighted mixture of components. That is, $o(x_v, y) = u^T f(x_v, y)$.
Our mixture model consists of a submodular facility location term and modular importance terms. The
facility location function is defined as $f_{fl}(X) = \sum_{v \in V} \max_{x \in X} sim(v, x)$ where $v$ is an element
from the ground set \( V \) and \( \text{sim}(v, x) \) measures the similarity between element \( v \) and element \( x \).

Facility-location thus models representativeness. The importance scores are taken from the VASNet model \([3]\) and the vsLSTM model \([30]\) trained on VISIOCITY. The weights of the model are learnt using the large margin framework as described in \([8]\) using many automatic ground truth summaries and a margin loss which combines the feedback from different evaluation measures. Specifically, given \( N \) pairs of a video and an automatic reference summary \((V, y_{gt})\), we learn the weight vector \( w \) by optimizing the following large-margin formulation \([26]\):

\[
\min_{w \geq 0} \frac{1}{N} \sum_{n=1}^{N} L_n(w) + \frac{\lambda}{2} ||w||^2,
\]

where \( L_n(w) \) is the generalized hinge loss of training example \( n \) and \( w \) is the weight vector. That is,

\[
L_n(w) = \max \left( w^T f(x_n^v, y) + l_n(y) - w^T f(x_n^v, y_{gt}) \right).
\]

This objective is chosen so that each ground truth summary scores higher than any other summary by some margin. For training example \( n \), the margin we chose is denoted by \( l_n(y) \) and is a linear combination of the normalized losses reported by our proposed measures. We call our proposed method VISIOCITY-SUM. We show that a simple model like this out-performs the current techniques (state of the art on TVSum and SumMe) on VISIOCITY dataset because of learning from multiple ground truth summaries and learning from multiple loss functions, each capturing different characteristics of a summary.

7 Experiments and Results

7.1 Implementation Details

For analysis of and comparison with human summaries, we generated 100 automatic summaries per video of about the same length as the human summaries. F1 score of any candidate summary is computed with respect to the human ground truth summaries following \([30]\). We report both avg and max. To calculate F1 scores of human summaries with respect to human summary, we compute max and avg in a leave-one-out fashion.

For analysis of and comparison of different techniques on the VISIOCITY dataset, we report their F1 scores computed against the automatically generated summaries as a proxy for human summaries. We generate 100 automatic summaries for each video. All target summaries are generated such that their lengths are 1% to 5% of the video length. We test the performance of three different representative state-of-the-art techniques on the VISIOCITY benchmark vsLSTM \([30]\) is a supervised technique that uses BiLSTM to learn the variable length context in predicting important scores. It learns from a combined ground truth in terms of aggregated scores. VASNet \([3]\) is a supervised technique based on a simple attention based network without computationally intensive LSTMs and BiLSTMs. It learns from a combined ground truth in terms of aggregated scores and outputs a predicted score for each frame in the video. DR-DSN \([32]\) is an unsupervised deep-reinforcement learning based model which learns from a combined diversity and representativeness reward on scores predicted by a BiLSTM decoder. It outputs predicted score for every frame of a video. To generate a candidate machine generated summary from the importance scores predicted by vsLSTM, VASNET and DR-DSN, we follow \([30]\) to convert them into machine generated summary of desired length (max 5% of original video). Our proposed model, VISIOCITY-SUM learns from multiple ground truth summaries and outputs a machine generated summary as a subset of snippets.

In all tables, AF1 refers to Avg F1 score, MF1 refers to Max F1 score (nearest neighbor score), IMP, MC, DT, DC and DSI refer to the importance score, mega-event continuity score, diversity-time score, diversity-concept score and diversity-similarity score respectively, as calculated by the proposed measures. All figures are in percentages.

7.2 Different human summaries have different characteristics

We asked a set of 11 users (different from the annotators) to create human summaries for two randomly sampled videos of each domain. The users were asked to look at the video without the audio and mark segments they feel should be included in the summary such that the length of the summary remains between 1% to 5% of the original video. The procedure followed was similar to that of SumMe \([6]\). We assess these human summaries qualitatively and quantitatively using the proposed set of performance measures and make the following observations. The human summaries are consistent with each other in as much as there are important scenes in the video, for example goals in Soccer videos. In the absence of such clear interesting events, the human summaries exhibit
<table>
<thead>
<tr>
<th>Domain</th>
<th>Technique</th>
<th>AF1</th>
<th>MF1</th>
<th>IMP</th>
<th>MC</th>
<th>DT</th>
<th>DC</th>
<th>DSI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Soccer</td>
<td>Human</td>
<td>30</td>
<td>45</td>
<td>56</td>
<td>55</td>
<td>75</td>
<td>84</td>
<td>85</td>
</tr>
<tr>
<td></td>
<td>Uniform</td>
<td>6</td>
<td>9</td>
<td>30</td>
<td>19</td>
<td>30</td>
<td>52</td>
<td>82</td>
</tr>
<tr>
<td></td>
<td>Random</td>
<td>5</td>
<td>9</td>
<td>30</td>
<td>22</td>
<td>30</td>
<td>51</td>
<td>81</td>
</tr>
<tr>
<td></td>
<td>Auto</td>
<td>27</td>
<td>37</td>
<td>83</td>
<td>82</td>
<td>82</td>
<td>90</td>
<td>80</td>
</tr>
<tr>
<td>Friends</td>
<td>Human</td>
<td>24</td>
<td>38</td>
<td>55</td>
<td>46</td>
<td>72</td>
<td>70</td>
<td>88</td>
</tr>
<tr>
<td></td>
<td>Uniform</td>
<td>5</td>
<td>9</td>
<td>31</td>
<td>7</td>
<td>89</td>
<td>66</td>
<td>90</td>
</tr>
<tr>
<td></td>
<td>Random</td>
<td>6</td>
<td>13</td>
<td>31</td>
<td>16</td>
<td>28</td>
<td>38</td>
<td>85</td>
</tr>
<tr>
<td></td>
<td>Auto</td>
<td>25</td>
<td>41</td>
<td>87</td>
<td>69</td>
<td>76</td>
<td>85</td>
<td>81</td>
</tr>
</tbody>
</table>

Table 4: Performance of Human and Auto summaries for different domains. TechTalk videos do not have MegaEvents.

more inconsistency with each other. A representative plot (for the scores of human summaries of "friends_5" video is presented in Figure 1). We note the following: a) proposed measures get good values on the human summaries as compared to uniform and random summaries, thus ascertaining their utility b) a human summary could score low on one measure and high on another measure c) the desired characteristics differ slightly across different domains (for example "importance" seems to be more important for soccer videos than diversity)

### 7.3 Automatically generated reference summaries are at par with human summaries

We compare automatically generated reference summaries with human summaries on our proposed measures and present the quantitative results in Table 4. For brevity, we present the numbers only for Soccer and Friends videos and provide numbers for other categories in the Supplementary. We see that automatic and human summaries are much better than random on all the evaluation criteria. Next, we see that both the human and the automatic summaries are close to each other in terms of the F1 metric. The automatic summaries have the highest Importance, Continuity and Diversity scores. This is not surprising as they are obtained in the first place by optimizing a combination of these criteria. Figure 2 shows a representat plot for min, mean, max of different measures for different summaries of soccer videos.

We also compare the human and automatic summaries qualitatively. We present some results in the project page. We see a considerable similarity in selections, though a perfect match of selections is neither possible nor expected, in keeping with the spirit of multiple correct answers. Some human summary videos and automatic ground truth summary videos are also reported at the project page. We see that a) it is very hard to distinguish the automatic summaries from human summaries and b) they form very good visual summaries in themselves.

![Figure 1: Different human summaries of same video perform differently on different measures](image1)

![Figure 2: Behavior of different measures for different types of summaries for Soccer videos. 'vis-cont' summaries are visually continuous summaries assembled by picking a set of continuous snippets](image2)
7.4 VISIOCITY Benchmark: Performance of different models on VISIOCITY

We test some state-of-the-art models and our simple enhancement VISIOCITY-SUM on VISIOCITY. We report the numbers for Soccer and Friends here and defer the rest to the Supplementary.

<table>
<thead>
<tr>
<th>Domain</th>
<th>Technique</th>
<th>AF1</th>
<th>MF1</th>
<th>IMP</th>
<th>MC</th>
<th>DT</th>
<th>DC</th>
<th>DSI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Soccer</td>
<td>Auto</td>
<td>59.3</td>
<td>93.3</td>
<td>83.2</td>
<td>84.3</td>
<td>82.6</td>
<td>85.9</td>
<td>76.2</td>
</tr>
<tr>
<td>DR-DSN</td>
<td></td>
<td>2.8</td>
<td>9.9</td>
<td>22.3</td>
<td>20.3</td>
<td>23.2</td>
<td>30.4</td>
<td>83.4</td>
</tr>
<tr>
<td>VASNET</td>
<td>28.4</td>
<td>43.4</td>
<td>63</td>
<td>49.3</td>
<td>62.1</td>
<td>67.4</td>
<td>75.2</td>
<td></td>
</tr>
<tr>
<td>vsLSTM</td>
<td>31.9</td>
<td>48.2</td>
<td>62.2</td>
<td>60.1</td>
<td>62</td>
<td>69.5</td>
<td>76.5</td>
<td></td>
</tr>
<tr>
<td>Ours</td>
<td>32.6</td>
<td>50.3</td>
<td>64.2</td>
<td>62.6</td>
<td>63.4</td>
<td>72.2</td>
<td>78.7</td>
<td></td>
</tr>
<tr>
<td>Random</td>
<td>3.4</td>
<td>9.3</td>
<td>25.7</td>
<td>18.5</td>
<td>25.5</td>
<td>39.2</td>
<td>80.5</td>
<td></td>
</tr>
<tr>
<td>Friends</td>
<td>AUTO</td>
<td>66.3</td>
<td>96.9</td>
<td>87.8</td>
<td>84.6</td>
<td>80.3</td>
<td>89.8</td>
<td>83.1</td>
</tr>
<tr>
<td>DR-DSN</td>
<td></td>
<td>4.3</td>
<td>9.4</td>
<td>19.1</td>
<td>6.9</td>
<td>65.7</td>
<td>51.5</td>
<td>98.5</td>
</tr>
<tr>
<td>VASNET</td>
<td>17</td>
<td>29.6</td>
<td>41</td>
<td>39.3</td>
<td>49</td>
<td>60.6</td>
<td>86.7</td>
<td></td>
</tr>
<tr>
<td>vsLSTM</td>
<td>15.5</td>
<td>27.2</td>
<td>40.4</td>
<td>39.2</td>
<td>64.7</td>
<td>59</td>
<td>91.1</td>
<td></td>
</tr>
<tr>
<td>Ours</td>
<td>17.4</td>
<td>31.2</td>
<td>42.5</td>
<td>40.5</td>
<td>50.2</td>
<td>64</td>
<td>90.3</td>
<td></td>
</tr>
<tr>
<td>Random</td>
<td>7.7</td>
<td>17.9</td>
<td>31.5</td>
<td>19.8</td>
<td>34.8</td>
<td>45.2</td>
<td>85.9</td>
<td></td>
</tr>
</tbody>
</table>

Table 5: Comparison of different techniques on VISIOCITY. TechTalk videos do not have MegaEvents

We make the following observations: a) DR-DSN tries to generate a summary which is diverse. As we can see in the results, it almost always gets high score on the diversity term. Please note that the way we have defined these diversity measures, diversity-concept (DC) and diversity-time (DT) have an element of importance in them also. On the other hand, diversity-sim (DSi) is a pure diversity term where DR-DSN almost always excels. b) Due to this nature of DR-DSN, when it comes to videos where the interestingness stands out and importance clearly plays a more important role, DR-DSN doesn’t perform well. In such scenarios, vsLSTM is seen to perform better, closely followed by VASNET. c) It is also interesting to note that while two techniques may yield similar scores on one measure, for example vsLSTM and VASNET for Soccer videos (Table 5), one of them, in this case vsLSTM, does better on mega-event continuity and produces a desirable characteristic in the summary. This further strengthens our claim of having a set of measures evaluating a technique or a summary rather than over dependence on one, which may not fully capture all desirable characteristics of good summaries. d) We also note that even though DR-DSN is an unsupervised technique, it is a state of the art technique when tested on tiny datasets like TVSum or SumMe, but when it comes to a large dataset like VISIOCITY, with more challenging videos, it doesn’t do well, especially on those domains where there are clearly identifiable important events for example in Soccer (goal, save, penalty etc.) and Birthday videos (cake cutting, etc.). In such cases, models like vsLSTM and VASNET perform better as they are geared towards learning importance. In contrast, since the interestingness level in videos like Surveillance and Friends is more spread out, DR-DSN does relatively well even without any supervision. e) VISIOCITY-SUM does better than all techniques on account of learning from individual ground truth summaries and a combination of loss functions.

8 Conclusion

In order to improve the objectivity and consistency in the design of video summarization benchmark datasets as well as their use in evaluating video summarization models, we present VISIOCITY, a large benchmarking dataset and demonstrated its effectiveness in real world setting. To the best of our knowledge, it is the first of its kind in the scale, diversity and rich concept annotations. We introduce a recipe to automatically create ground truth summaries typically needed by the supervised techniques. Motivated by the fact that different good summaries have different characteristics and are not necessarily better or worse than the other, we propose an evaluation framework better geared at modeling human judgment through a suite of measures than having to overly depend on one measure. Finally we report the strengths and weaknesses of some representative state of the art techniques when tested on this new benchmark and demonstrate the effectiveness of our simple extension to a mixture model making use of individual ground truth summaries and a combination of loss functions. We hope our attempt to address the multiple issues currently surrounding video summarization as highlighted in this work, will help the community advance the state of the art in video summarization. We make VISIOCITY available through the project page at https://visiocity.github.io/ and invite the researchers to test their algorithms on VISIOCITY benchmark.
References


Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [No] Since this is a dataset paper, there are no apparent limitations. Any video summarization technique can be tested against this dataset.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] We do not foresee any potential negative societal impacts of this work.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]

3. If you ran experiments (e.g. for benchmarks)...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] All assets are available for download from the project page
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] In the implementation details section.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] We have reported average results across random runs wherever applicable and mentioned in the implementation details
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] All experiments were run on an nVidia RTX 2080Ti GPU.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] We download the publicly available videos from YouTube
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] All assets are available to download from the project page
   (d) Did you discuss whether and how consent was obtained from people whose data you’re using/curating? [N/A] We have created our own dataset
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] All videos used are publicly available and downloaded from YouTube. For surveillance videos with personally identifiable information, prior permission was sought from the participants

5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [Yes] Included in the Supplementary
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [Yes] Included in the Supplementary