IET Journals
The Institution of Engineering and Technology

Video Summarization by Deep Visual and Categorical Diversity

ISSN 1751-8644
doi: 0000000000
www.ietdl.org

Pedro Atencio[1], *German Sanchez*[1], *John Branch*[2], *Claudio Delrieux*

[1] *Proffesor: Faculty of Engineering, Instituto Tecnologico Metropolitano, Medellin, Colombia*
[2] *Proffesor: Faculty of Engineering, Universidad del Magdalena, Santa Marta, Colombia*
[3] *Proffesor: Faculty of Mines, Universidad Nacional de Colombia, Medellin, Colombia*
[4] *Proffesor: Computing Department, Universidad Nacional del Sur, Bahia Blanca, Argentina*
** E-mail: pedroatencio@itm.edu.co*

Abstract: We propose a video summarization method based on visual and categorical diversity by transfer learning. Our method extracts visual and categorical features from a pre-trained deep convolutional network (DCN) and a pre-trained word embedding matrix. Using visual and categorical information we obtain video diversity, which it is used as an importance score to select segments from the input video that best describes it. Our method also allows to perform queries during the search process, in this way personalizing the resulting video summaries according to the particular intended purposes. The performance of the method is evaluated using different pre-trained DCN models in order to select the architecture with the best throughput. We then compare it with other state-of-art proposals in video summarization using a data-driven approach with the public dataset `SumMe`, which contains annotated videos with per-fragment importance. The results show that our method outperforms other proposals in most of the examples. As an additional advantage our method requires a simple and direct implementation that does not require a training stage.

1 Introduction and Previous Work

Digital video is a widespread medium in several contexts and applications. The generation and availability of digital video from different sources is growing at an exponential rate. This poses several challenges for users that require to retrieve information in vast collections of videos. One major way to simplify and accelerate the access to a particular information item in a video sequence is to provide abridged (albeit complete in some sense) representations of the whole content. This significantly reduces the burden of having to watch complete videos to decide whether and where the required information is present. Video summarization (VSUM) aims to provide these condensed versions in a consistent and predictable way.

Summarization techniques must produce an intelligible output that can be useful to human users. There are multiple aspects to consider in such kind of management of digital video. On one hand, any processing task must consider the capture, encoding, and compression techniques that are applied in digital media. On the other hand, semantic and psychological features should be taken into account for an adequate processing and manipulation. Semantics is mainly expressed in words [1], for which it is necessary to represent video contents both regarding visual and linguistic information. However, simple language-based techniques as tagging and string search are too superficial to provide useful results. Even more, in query-based VSUM, when users request to recall whether and where a given object or action is present in a video set, it is necessary to take into account deeper linguistic information, *i.e.*, queries and their linguistic space must be somehow combined with visual information to make sense. Thus, multi-modal information representation techniques are indispensable for query-based VSUM.

VSUM is commonly treated as a regression or ranking problem where some features are extracted from video-frames and used as inputs, and a set of key-frames or user-annotated scores [2–7] as outputs. Earlier approaches focused exclusively in supervised visual features extracted from video [6, 8, 9] as SIFT, HoG or optical flow, among others. Lee et. al. in [10] used features as eye fixation, object frequency and interaction to predict scene importance. Depending of the nature of the video or the search task, domain-specific features may be an aid in VSUM. For example, game-specific rules for sport video analysis was proposed by Shih et. al. in [11], actor recognition [12], and subtitle analysis [13] for movie summarization. Egocentric video analysis lately emerged as another significant VSUM context, because of its characteristic high volume and diversity, which has motivated researchers to propose general VSUM methods that could complement visual information with associated annotations and meta-data. For instance, `SumMe` [2] and `TVSUM` [14] are two widely used VSUM datasets with egocentric videos, both with human-annotated score of importance per video segment.

Recently, deep learning has been applied to VSUM from multiple approaches. Otani et al. [5] proposed a method to train a coordinated representation space from a video-to-text (VTT) dataset and posteriorly used it to generate a regression model for VSUM. Temporal analysis using long-short term memory networks (LSTM) and transfer learning from DCN (Deep Convolutional Network) was used in [15]. Generative adversarial networks (GANs) was used in [16] to formulate the VSUM problem as a generator/discriminator challenge, where generator select best frames (summary) from input video and reconstruct it from those video-frames, then discriminator compares input-video and reconstructed video as a classification problem. For this purpose, an architecture based on DCN and LSTM was constructed.

It is possible to approximate frame importance by the uniqueness or diversity of a group of frames. Uniqueness is related with the dissimilarity or difference of descriptors for consecutive frames [17]. Classical approaches consider a processing pipeline where video frames are first pre-processed to improve quality, after which they are represented using a static set of descriptors [18]. These descriptors are mainly low-level visual features, i.e., color, textures, histograms, among others. Finally, a supervised criterion is designed using specific descriptors to estimate an importance score that allows selecting frames for the resulting summarized video. This approximation has some limitations that impairs the possibility of using multi-modal information. In particular, the use of hand-crafted descriptors, and also the importance criterion, have a high impact in the resulting summary, due to their *ad-hoc* nature.

Human attention is another information source closely related to diversity and commonly used for video summarization. Visual-auditive saliency and attention have been explored for VSUM task in [19, 20]. Varini et. al. in [21], used a combination of HMM and diversity from Bag of Words difference between consecutive segments to create a video summary. Gygli et. al. in [2] represented attention for video summarization as a nonlinear combination of spatial features and temporal saliency expressed as temporal differences.

Nevertheless, many claim that frame importance cannot be done entirely without previously known context information, including recording purpose, user preferences, and overall history contained in the video, among other things [2]. Personalization is one the most recent topics of interest in video summarization because different users will summarize a video differently based on their specific interests. Initial approximations explored this venue by capturing more inputs from the user while doing this task, for example, gaze-tracking [22], BCI devices [23, 24] or states of attention [21]. These works have the limitation that user intention is not known previous to the video summarization task, and also require extra equipment.

Recent work focused on introducing queries in natural languages during the VSUM process, as a means to specify a specific purpose. These queries may be expressed either as a vector of words of interest, which can be a sentence in natural language [25] or as a set of categorical terms (objects of interest) [26–28]. The first approximation requires NLP techniques to transform an arbitrary sentence into a manageable structure. The second approximation represent queries as a vector of words related to the objects of interest for the user. However, as far as a thorough exploration of recent advances in VSUM may reveal, diversity from combined deep visual and categorical features has not been previously used. Also, the possibility to deliver personalized video summary guided by user query is an important feature that is certainly sought for in popular video repositories like YouTube and others. For these reasons we propose a novel VSUM method based on a categorical diversity estimation found combining both visual features and semantic categories inferred by user queries. The direct nature of our method without the requirement of a training stage allows rapid adoption and implementation for industrial and commercial applications.

2 Materials and Methods

Our method uses a pre-trained DCN architecture as a general visual descriptor for frames in V, extracting deep-features from internal layers [29]. Then, using the last layer of the DCN we can obtain a word set related to detected categories that appear in each video frame V_i. We generate a categorical representation of this word set using a pre-trained word embedding. Finally, a linear criteria of mutual penalization of visual and categorical representation is constructed. A schema of the proposed method is shown in Fig. 1.

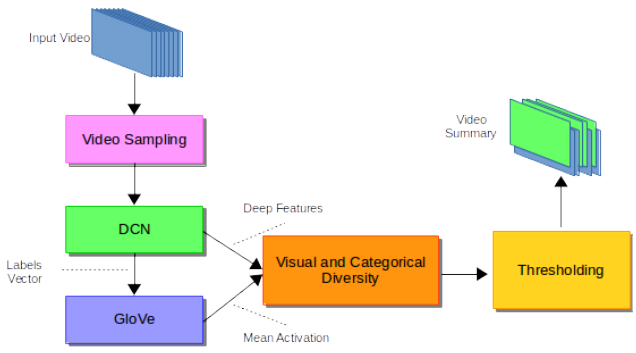

Fig. 1: Schema of the proposed method.

2.1 Visual Representation

As mentioned in [29], using the right set of features, almost any AI problem can be solved. In particular, visual data representation is a complex problem due to the non-structured high-dimensional nature of images. In this context, Deep learning may be used not only as a black box that maps input (stimulus) to output (response), but also to discover the best input representation, a task called *representation learning* or deep features, which basically consists in using a hidden layer of a pre-trained network model M as a general representation or descriptor of an input data.

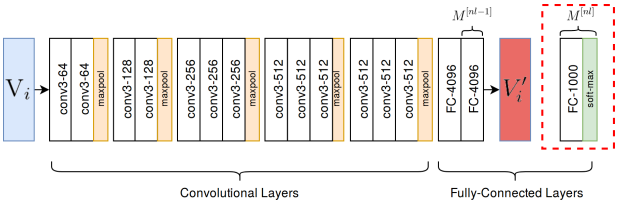

Fig. 2: VGG16 architecture. We compute the $M^{[nl-1]}$ activation as a descriptor of input frames V_i.

For visual data representation, it is common to use DCN architectures previously trained for a classification task. The main idea behind this approach is that it is possible to transfer learned knowledge from a previous task to accelerate training for a new task. We use a pre-trained DCN model M using image-net dataset [30], as feature extractor for frames V_i. For example, VGG-16 [31] model has 16 layers and approximately 130 M parameters. The penultimate layer, $M^{[nl-1]}$, *i.e.*, the layer before softmax 1000-dimensional probabilities of image-net categories, is 4096-dimensional. Then, we describe every frame V_i of input video V as follows:

$$V'_i = M^{[nl-1]}(V_i)$$

That is, we compute penultimate layer activation for a DCN model M with nl layers, given a video-frame V_i. For this, we remove the layer of DCN model and compute a feed-forward propagation through the network. A graphical depiction of this representation scheme is shown in Fig.2.

2.2 Categorical representation

The complete forward propagation of a pre-trained DCN model generates an activation from the last layer $M^{[nl]}(V_i)$ given an input video frame V_i. This activation consists of a 1000-dimensional vector with the probabilities (softmax output) for every ImageNet category to appear in V_i. Using this vector we select the k category indices with the highest probability to appear in V_i. These indices are represented as an array γ which it is transformed to a one-hot encoding array t of size $[nd, 1]$ using word-embedding matrix vocabulary, where nd is the dimension of the embedding representation. Finally, a dot product between word-embedding matrix B and t is computed to obtain an embedded representation E_j for a visual category j from image-net.

Due to the fact that each category j in video frame V_i is associated with a probability from the softmax layer $M^{[nl]}$, we compute a weighted average representation G_i as in Oosterhuis et. al. in [28]. The following equation describes the weighted average used for the word embedding representation G_i for video-frame V_i:

$$G_i = \frac{1}{k} \sum_{j}^{k} \psi_j E_j, \tag{1}$$

with $\psi_j = M_j^{[nl]}(V_i)$. Notice that categories probabilities ψ are obtained from the last layer from DCN model M. A complete sequence to obtain categorical representation G_i is presented in Algorithm 1.

Algorithm 1: Categorical representation of V_i using embedding matrix.

Input : video frame V_i, embedding-matrix B, DCN model M, k first visual categories
Output: weighted average word-embedding activation G_i

begin
  $\gamma = argsort(M^{[nl]}(V_i))[0:k]$;
  $E = zeros([nd, 1])$;
  for $j \leftarrow 0$ **to** $k-1$ **do**
    $t = zeros([nd, 1])$;
    $t[\gamma[j]] = 1$; // one-hot encoding
    $E_j = B \cdot t$; // embedding representation
    $\psi_j = M_j^{[nl]}(V_i)$; // weight from softmax layer
    $G_i = G_i + \psi_j E_j$; // weighted sum
  end
  $G_i = \frac{1}{k} G_i$; // weighted average
  return G_i;
end

2.3 Visual and categorical diversity

Using the concept of diversity expressed as temporal differences, we obtain visual diversity D_i for an input video-frame V_i as follows:

$$D_i = \left\| \frac{dV_i'}{dt} \right\| = \left\| \frac{V_{i+\Delta t}' - V_i'}{\Delta t} \right\|,$$

where V_i' is the deep-feature vector extracted from the penultimate layer of DCN model as explained previously, and Δt is 1 because input layer is previously sampled to $\frac{1 frame}{second}$. Notice that we compute the L2-norm of the derivative in order to obtain a scalar-valued visual diversity. Using same approach as visual diversity D_i we compute categorical diversity K_i as temporal differences from categorical representations G_i:

$$K_i = \left\| \frac{dG_i}{dt} \right\| = \left\| \frac{G_{i+\Delta t} - G_i}{\Delta t} \right\|.$$

Finally, we scale D_i and K_i in the range $[0, 1]$ in order to avoid magnitudes differences from visual and categorical representations V_i and G_i.

2.4 Combined visual and categorical diversities

Visual and categorical diversity, D_i and K_i respectively, for a video frame V_i are related but provide measures in different description domains for the frame. In other words, for a given video frame V_i we can obtain consensual diversity description (or lack thereof), or a high visual diversity and low categorical diversity (or the other way around). Then, we can assume that we need to have both high visual diversity D_i and high categorical diversity K_i to consider a frame to have a high diversity ϑ. In other words, if a video frame is visually diverse but not categorically, or viceversa, then it is not regarded as important. From this analysis, we balance visual and categorical diversity using a linear relation using coefficients c_0 and c_1. In Eq. 2, diversity ϑ_i expressed in terms of D_i and K_i, with the special case $c_0 = c_1 = 0.5$ as the mean of visual and categorical diversity.

$$\vartheta_i = c_0 D_i + c_1 K_i \tag{2}$$

Finally, we apply a Gaussian smoothing to the time series generated for the sequence of ϑ_i in order to generate a more continuous diversity function along video frames. This allows to generate video summaries with soft transitions between segments. In Fig. 3 it is shown the combined diversity ϑ with respect to D_i and K_i for video *St. Marteen Landing* from database SumMe. As explained previously, low values for D_i or K_i yield low diversity values, as happens

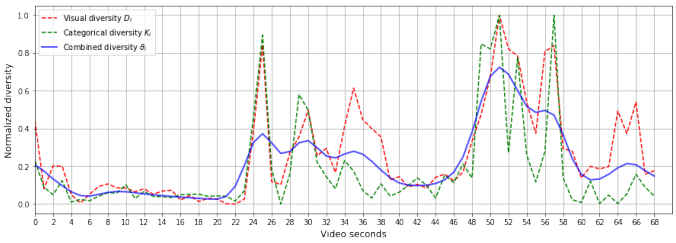

Fig. 3: Gaussian smoothed (σ =1.5) and normalized combined diversity ϑ_i with respect to D_i and K_i for video *St. Marteen Landing* from database SumMe [2].

for example in seconds 44 and 52. On the contrary, when both values are hight, as in seconds 25 and 56, the corresponding frames are considered important.

2.5 Summary generation

For this purpose, we apply thresholding to ϑ as described in the following equation:

$$\begin{aligned} S(\alpha) &= V_i \mid \vartheta_i > \tau \\ &\text{restricted to:} \\ (|S| - \alpha\,|V|) &\rightarrow 0, \end{aligned} \tag{3}$$

where $|S|$ and $|V|$ are the lengths of S and V respectively and α is a summary length scalar, usually $\alpha = 0.15$. Then, we need to find a value τ such that the number of all frames with diversity ϑ_j greater than τ, is closely to $\alpha|V|$.

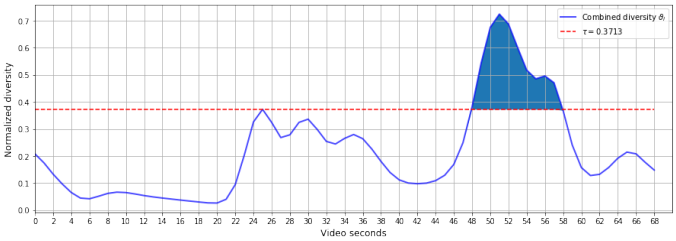

Fig. 4: Summary generation from ϑ_i for video *St. Marteen Landing* from database SumMe [2], using the proposed method. Frames in the blue region will be used as a summary for the input video.

The summary generation process is shown in Fig.4. Notice that dashed red line illustrate the value of τ which was found using an iterative process as explained previously. The blue shaded region illustrates the subset of frames (15% of input video length $|V|$) that constitutes the resulting summary.

2.6 Query injection

We represent an user query as a vector $q = \{w_0, w_1, w_2, ..., w_n\}$ of words w. In order to avoid direct-match between q and visual categories j detected by DCN model M, each word of the query is mapped to a vector space using a word-embedding matrix E as explained in previous sections, in a similar manner as in Oosterhuis et al. [28]. We assume that all words in the query have the same importance for the user, as opposed to what is proposed in [28], so we do not weight words w in the query q. Then, we calculate the average of word-vectors for k words in query as follows:

$$G_{query} = \frac{1}{k}\sum_{j}^{k} E_j \tag{4}$$

The categorical representation G_i (section 2.2) extracted from video-frame V_i, allow us to compare categorical similarity H_i of q and V_i in a direct manner using the cosine similarity as follows:

$$\begin{aligned} cos(w_i, w_j) &= \frac{w_i \cdot w_j}{\|w_i\| \, \|w_j\|} \\ H_i &= cos(G_i, G_{query}) \end{aligned} \tag{5}$$

Finally, we can represent query-combined diversity φ as a linear relation between combined diversity ϑ_i and query similarity S_i as follows:

$$\begin{aligned} \varphi_i &= c_0\vartheta_i + c_1 H_i = \\ \varphi_i &= c_0 D_i + c_1 K_i + c_2 H_i \end{aligned} \tag{6}$$

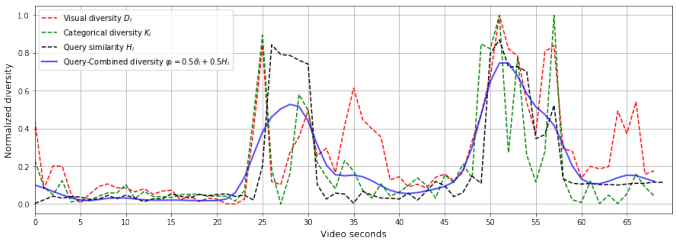

Fig. 5: Video diversity biased by query similarity for $q = \{airship, aircraft\}$ over video *St. Maarten Landing* from dataset SumMe. Query similarity in black, visual diversity in red and categorical diversity in green.

In Fig. 5 it is shown an example of query injection to find combined diversity ϑ using Eqs. 4, 5, and 6.

3 Experiments

In the following experiments, we first compare visual and categorical diversity from different combinations of DCN models in order to determine if there exist significant differences between them. Then, each combination is evaluated using SumMe dataset and performance criteria defined later in this chapter, and we select the model with the best performance. Finally, we compare the best model with respect to state-of-the-art works.

3.1 Data

SumMe: Video to user-importance dataset. Proposed and published by Gygli et. al. in [2], this dataset contains 25 videos, organized by three categories: *egocentric*, *moving* and *static*. For each video, scores per segments were manually annotated by different individuals. This score is related with an importance scale in the range $[0, 1]$, (see Figure 6) for minimum and the maximum degree of importance for the user, respectively. Authors report that this was the first dataset with segments annotations rather than key-frames. This dataset has been widely used for different video summarization approaches on literature, allowing us to compare with respect to other authors.

Video Preprocessing. Input video V is uniformly sampled to one frame per second as similar to related works [1, 5, 32]. We use this approach in order to reduce computational processing and due to the need that our model must be as simple as possible, avoiding alternatives like clustering or super-frame segmentation [2]. Videos from SumMe dataset mostly have a 30 fps rate. We consider that at this frame rate will not occur a significant visual and categorical diversity in less than a second. Then, we take a frame per second in order to reduce processing time, that is, $|V_{sampled}| = \frac{|V|}{fps}$.

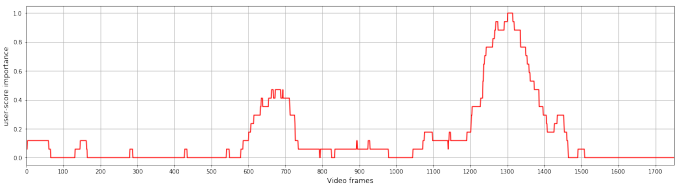

Fig. 6: Averaged user-importance score for video *St. Marteen Landing* from database SumMe [2].

3.2 Performance criteria

Evaluation of a generated summary using SumMe dataset, consists in performing a measure of accuracy between video summaries extracted from a computational method and video summaries extracted from estimated importance given by N human users. For this purpose, Gygli et. al. [2] proposed the use of pair-wise *f-measure* evaluation metric between a generated summary $S(\alpha)$ and human-made scores $U(\alpha)$ as expressed in Equation 7.

$$F(S, U, \alpha) = \frac{1}{N-1}\sum_{j=1}^{N} 2\frac{p_{(S(\alpha),U_j(\alpha))} r_{(S(\alpha),U_j(\alpha))}}{p_{(S(\alpha),U_j(\alpha))} + r_{(S(\alpha),U_j(\alpha))}} \tag{7}$$

Where $p_{(S(\alpha),U_j(\alpha))}$ and $r_{(S(\alpha),U_j(\alpha)}$ are precision and recall between generated summary and interestingness score made by user j. Notice that, final score is the averaged result of the summary S with respect to each user annotation U_j at $\alpha|V|$ length of the original video.

3.3 Baselines

We evaluate our method with respect to the following approximations:

- **Random sampling**: Video summarization by taking random frames is commonly used a base comparison on literature due to the fact that any proposed video summarization method must be superior to this approach.
- **Interestingness-based[2]**: Original work by Gygli et. al., where it is proposed SumMe dataset and a video summarization method based on a regression model that uses a combination of features related with frame-interestingness: attention, aesthetics, presence of landmarks, faces and object tracking, to predict per-frame importance.
- **Deep semantic features [5]**: As proposed by Otani et. al., this method uses coordinated representations which authors refers as semantic features, trained over a video-to-text dataset. This representations are then used in a regression model to predict per-frame importance.

3.4 Results

3.4.1 Visual and categorical diversity relationship: As explained in previous sections, our method lies in the use of a pretrained DCN and word-embedding models for extracting visual and categorical diversity respectively. Visual diversity D depends exclusively of activations from DCN model. Categorical diversity K depends of DCN model and word-embedding activations. In this order of ideas, it is possible to ask the following questions:

- Will different DCN models generate different/similar visual diversity D?
- Will Different DCN models generate different/similar categorical diversity K?
- Will visual and categorical diversity D and K, be related for each DCN model?

To answer these questions we measured correlation coefficients for visual and categorical diversity through each video in dataset SumMe, using different DCN models. In tables 1 and 2 can be observed the mean correlation for D and K respectively.

Notice in table 1 that, in general, correlation coefficients for D are greater than 0.7, which can be interpreted as that visual diversity is strongly related across different DCN models. In other words, a similar visual diversity is expected to be obtained when using any of the evaluated DCN models.

It is also possible to observe that models of the same family such as VGG16/VGG19 and DenseNet/ResNet50 will have a much stronger correlation for visual diversity.

Correlation coefficients for categorical diversity K can be observed in 2. Notice that, in general, categorical diversity correlation across DCN models is below 0.4, which can be interpreted as a weak relation between models. In other words, the selection of a particular DCN model will result in a different categorical diversity. Similarly to visual diversity, models of the same family such as VGG16/VGG19 and DenseNet/ResNet50 will have a strong correlation for categorical diversity.

Finally, we obtained correlation coefficients between D and K for each evaluated DCN model, as can be observed in table 3. For any DCN model, Visual and categorical diversity have a moderate positive relationship, as correlation coefficients are 0.5 approximately. In other words, as explained in previous sections, although D and K are related and depends of a DCN model, we can expect that for a given input video-frame V_i we can obtain a high value of D_i and low K_i or the contrary. Thus, we can expect the use of combined diversity ϑ in a video summarization task, will generate a similar response to human users.

3.4.2 Evaluation between DCN models: We evaluated performance of our model using different pretrained DCN models over SumMe dataset videos. Evaluation was made using f-score as presented in equation 7.

In table 6 (see section 6) it can be observed the score by video in SumMe dataset, using our method with different popular DCN pretrained models. InceptionV3 and InceptionResNetV2 models obtains the highest f-score (bold) for most videos (6 videos each model). In terms of mean or averaged score, **InceptionV3** is the model with highest f-score and will be used as a base DCN for comparison with state-of-art works. It is important to mention that although InceptionV3 is the best DCN model in general terms, mean f-score for each evaluated DCN model are similar with 0.189 as lowest f-score and 0.209 as highest f-score.

3.4.3 State-of-art comparison: In table 7 (see section 6) it is shown the results of quantitative evaluation for our method and different computational methods as explained in section 3.3. We shown scores for human annotators as reported by author in [2] as follows:

- **Minimum score (Min)**: Lowest score of all human annotators with respect to mean score.
- **Mean score (Mean)**: Average of scores made by each human annotators with respect to others. For example, if a video is annotated by 20 users, then f-score is computed for each annotator with respect to the others (19). Finally all previous are averaged.
- **Maximum score (Max)**: Highest score of all human annotators with respect to mean score.

For computational methods, we report best scores per video in bold and performance relative to human annotators.

Our method, obtains higher performance than method proposed by Otani et. al., with 67% and 59% respectively, relative to human average performance. Method proposed by Gygli et. al. still obtains the highest mean f-score with a performance with respect to human average of 75%. Also we obtained bests scores in $\frac{9}{25}$ videos, Gygli in $\frac{12}{25}$ videos, and Otani in $\frac{4}{25}$ videos.

Performance by video category is presented in table 4. Notice that our method obtains the highest score for *Static* category with a remarkable difference with respect to the other computational approaches. *Moving* category represents the lowest score for our method and the higher difference in performance with respect to method proposed by Gygli. Finally, in *Egocentric* category we obtain a similar performance to last model.

We consider important to mention that our method is simpler than computational methods proposed by Gygli and Otani, in terms that relies in using a single DCN pretrained model and word-embeddings which do not require training stage.

3.4.4 Error analysis: Our method got low scores when videos contain complex stories in terms of actions and interactions where there is not high diversity in visual properties of V_i or objects on scene, in other words, when video importance is not related with diversity but the story. Examples of this kind of videos are *playing_ball*, *Excavators River Crossing* and *Playing on Water Slide*, where our method obtains its lowest scores.

Table 1 Correlation coefficients for visual diversity D using different DCN models.

Model (Visual Diversity)	VGG16	VGG19	Xception	InceptionV3	ResNet50	InceptionResNetV2	DenseNet
VGG16	1,0	0,931	0,802	0,756	0,871	0,706	0,859
VGG19	0,931	1,0	0,797	0,754	0,864	0,705	0,862
Xception	0,802	0,797	1,0	0,776	0,811	0,753	0,816
InceptionV3	0,756	0,754	0,776	1,0	0,782	0,749	0,783
ResNet50	0,871	0,864	0,811	0,782	1,0	0,728	0,882
InceptionResNetV2	0,706	0,705	0,753	0,749	0,728	1,0	0,748
DenseNet	0,859	0,862	0,816	0,783	0,882	0,748	1,0

Table 2 Correlation coefficients for categorical diversity K using different DCN models.

Model (Categorical Diversity)	VGG16	VGG19	Xception	InceptionV3	ResNet50	InceptionResNetV2	DenseNet
VGG16	1,0	0,527	0,295	0,225	0,332	0,246	0,301
VGG19	0,527	1,0	0,289	0,262	0,351	0,280	0,342
Xception	0,295	0,289	1,0	0,280	0,284	0,266	0,327
InceptionV3	0,225	0,262	0,280	1,0	0,238	0,300	0,276
ResNet50	0,332	0,351	0,284	0,238	1,0	0,234	0,369
InceptionResNetV2	0,246	0,280	0,266	0,300	0,234	1,0	0,293
DenseNet	0,301	0,342	0,327	0,276	0,369	0,293	1,0

In figure 7 (see section 6) we show examples of videos from dataset SumMe, with associated human scores of importance (red) and summary made by our method (blue). In video *playing_ball* main visual objects or categories are ball, dog and bird. These objects are always present of video-frames, so it is expected a low categorical diversity. Moreover, although video is moving, there are not important transitions or scene changes that generate high visual diversity. In this case, importance as annotated by users, is related with interactions between *bird*, *dog* and *ball*.

It is also important to consider that due to the fact that our method relies on a pre-trained DCN model, performance of summarization is related with performance of our DCN model, i.e., if DCN model does not correctly detect visual categories on video frame, diversity will be affected. Possible approximations to solve this limitation with the construction of joined-representations that can map time-related phenomena like actions or interactions.

On the contrary, in videos like *car_over_camera* and *St. Maarten Landing* (see figure 7 on section 6), importance is highly related to visual differences and changes of objects in scenes, which our method is able to capture as visual and categorical diversity. In these cases, we obtained higher performance than other computational methods.

In table 5 are shown query examples for three video on dataset SumMe. Videos *Playing Ball* and *Playing on Water Slide* are examples videos where we obtained low scores using our method. Notice that, injecting specific queries we were able to improve f-score. Also, using queries, was possible to improve result for videos that obtained high scores using global method. As was previously mentioned, due to the static nature of our method, query-injection will not be able to improve results on cases where video importance is related with actions or interactions which can not be detected by DCN models and therefore can not be represented by queries of categories.

4 Conclusions

We have developed a model based on visual and categorical diversity using pre-trained DCN models and word-embeddings. The main hypothesis is that importance of a video segment is related with diversity, i.e., the less diverse a segment, the less important. Literature explored visual diversity in previous works, but we have extended to a visual and categorical diversity. Experiments show that it is possible to use this combined diversity for a video summarization task. Although the simplicity of constructed model in terms of its architecture and the linear relation of visual and categorical diversity that was used, performance (f-score) is close or superior to state-of-art works. This motivates us to continue exploring in this direction of research. Some conclusions we can made based on experiments and results are: **a)** Visual diversity D, obtained from different DCN models is highly correlated, i.e. we can obtain a similar visual representation using any pre-trained DCN model, **b)** categorical diversity K, obtained from different DCN models and a word-embedding (GloVe) presents a low correlation, i.e. categorical representation vary significantly for different DCN models, **c)** although K depends of D, both representations are not highly related, i.e., a video-frame can be highly visually diverse but not categorically, or the contrary, **d)** we obtained best mean performance (f-score: 0.209) using InceptionV3 as pre-trained DCN model and **e)** injection of user queries can be used along with video diversity, to personalize summaries and specify user intentions previous to VSUM task.

As future work we will explore representing sequential nature of video to improve performance in videos where importance is related with actions and interactions between elements on scene and and the use of Region-based DCN models (R-CNN) to represent spatial interactions between scene objects.

Finally, simplicity and direct nature of our method allows to use it without a training stage, which is important for rapid implementation and processing for industrial applications.

Table 3 Correlation coefficients between visual diversity D and categorical diversity K using different DCN models.

DCN Model	correlation(D, K)
VGG16	0,554
VGG19	0,555
Xception	0,512
InceptionV3	0,504
ResNet50	0,471
InceptionResNetV2	0,587
DenseNet	0,506

Table 4 Mean f-measure per video category for different computational methods
(higher is better). For each video category, best result is shown in bold.

Category	Random	Gygli [2]	Otani [5]	Ours (DCN: Inception V3)
Egocentric	0,140	**0,226**	0,181	0,216
Moving	0.138	**0,225**	0,182	0,176
Static	0.138	0,285	0,186	**0,342**

5 References

1 S. Yeung, A. Fathi, and L. Fei-Fei, "VideoSET: Video Summary Evaluation through Text," *arXiv preprint arXiv:1406.5824*, 2014.
2 M. Gygli, H. Grabner, H. Riemenschneider, and L. Van Gool, "Creating Summaries from User Videos," in *ECCV*, 2014.
3 P. Varini, G. Serra, and R. Cucchiara, "Personalized Egocentric Video Summarization of Cultural Tour on User Preferences Input," *IEEE Transactions on Multimedia*, vol. PP, no. 99, p. 1, 2017.
4 K. Zhang, W.-L. Chao, F. Sha, and K. Grauman, "Summary Transfer: Exemplar-based Subset Selection for Video Summarization," *CoRR*, vol. abs/1603.0, 2016.
5 M. Otani, Y. Nakashima, E. Rahtu, J. Heikkilä, and N. Yokoya, "Video Summarization using Deep Semantic Features," *CoRR*, vol. abs/1609.0, 2016.
6 B. Gong, W.-L. Chao, K. Grauman, and F. Sha, "Diverse Sequential Subset Selection for Supervised Video Summarization," in *Proceedings of the 27th International Conference on Neural Information Processing Systems*, NIPS'14, (Cambridge, MA, USA), pp. 2069–2077, MIT Press, 2014.
7 M. Gygli, H. Grabner, and L. V. Gool, "Video summarization by learning submodular mixtures of objectives," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3090–3098, 2015.
8 Y. Gong and X. Liu, "Video Summarization and Retrieval Using Singular Value Decomposition," *Multimedia Syst.*, vol. 9, no. 2, pp. 157–168, 2003.
9 J. Iparraguirre and C. Delrieux, "Online Video Summarization Based on Local Features," *International Journal of Multimedia Data Engineering and Management (IJMDEM)*, vol. 5, pp. 41–53, 2014.
10 Y. J. Lee, J. Ghosh, and K. Grauman, "Discovering important people and objects for egocentric video summarization," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1346–1353, 2012.
11 H. C. Shih, "A Survey on Content-aware Video Analysis for Sports," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. PP, no. 99, p. 1, 2017.
12 J. Sang and C. Xu, "Character-based Movie Summarization," in *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, (New York, NY, USA), pp. 855–858, ACM, 2010.
13 N. Zlatintsi, P. Maragos, A. Potamianos, and G. Evangelopoulos, "A saliency-based approach to audio event detection and summarization," 2012.

Table 5 Query-injection examples for three videos on dataset SumMe. For each video it is shown score using combined diversity ϑ, query vector q and score using query combined diversity φ.

Videoname	Query q	Global VSUM ϑ	Query-based VSUM φ
Playing Ball	{german, shepperd, ball}	0,097	0,281
Playing on Water Slide	{water, kids}	0,074	0,183
St. Maarten Landing	{airship, aircraft}	0,469	0,553

14 Y. Song, J. Vallmitjana, A. Stent, and A. Jaimes, "TVSum: Summarizing web videos using titles," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5179–5187, 2015.
15 K. Zhang, W.-L. Chao, F. Sha, and K. Grauman, "Video Summarization with Long Short-term Memory," *ECCV*, pp. 1–24, 2016.
16 B. Mahasseni, M. Lam, and S. Todorovic, "Unsupervised video summarization with adversarial LSTM networks," in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017.
17 A. G. del Molino, C. Tan, J. H. Lim, and A. H. Tan, "Summarization of Egocentric Videos: A Comprehensive Survey," *IEEE Transactions on Human-Machine Systems*, vol. 47, pp. 65–76, 2 2017.
18 D. Borth, T. Chen, R. Ji, and S.-F. Chang, "SentiBank: Large-scale Ontology and Classifiers for Detecting Sentiment and Emotions in Visual Content," in *Proceedings of the 21st ACM International Conference on Multimedia*, MM '13, (New York, NY, USA), pp. 459–460, ACM, 2013.
19 Y.-F. Ma, L. Lu, H.-J. Zhang, and M. Li, "A User Attention Model for Video Summarization," in *Proceedings of the Tenth ACM International Conference on Multimedia*, MULTIMEDIA '02, (New York, NY, USA), pp. 533–542, ACM, 2002.
20 M. Sun, A. Farhadi, B. Taskar, and S. Seitz, "Summarizing Unconstrained Videos Using Salient Montages," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 11, pp. 2256–2269, 2017.
21 P. Varini, G. Serra, and R. Cucchiara, "Personalized Egocentric Video Summarization for Cultural Experience," *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval - ICMR '15*, vol. PP, no. 99, pp. 539–542, 2015.
22 J. Xu, L. Mukherjee, Y. Li, J. Warner, J. M. Rehg, and V. Singh, "Gaze-enabled egocentric video summarization via constrained submodular maximization," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2235–2244, 2015.
23 H. W. Ng, Y. Sawahata, and K. Aizawa, "Summarization of wearable videos using support vector machine," in *Proceedings. IEEE International Conference on Multimedia and Expo*, vol. 1, pp. 325–328, 2002.
24 K. Aizawa, K. Ishijima, and M. Shiina, "Summarizing wearable video," in *Proceedings 2001 International Conference on Image Processing (Cat. No.01CH37205)*, vol. 3, pp. 398–401, 2001.
25 B. Xiong, G. Kim, and L. Sigal, "Storyline representation of egocentric videos with an applications to story-based search," *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2015 Inter, pp. 4525–4533, 2015.
26 A. Sharghi, J. S. Laurel, and B. Gong, "Query-Focused Video Summarization: Dataset, Evaluation, and A Memory Network Based Approach," 7 2017.
27 A. Sharghi, B. Gong, and M. Shah, "Query-Focused Extractive Video Summarization," *CoRR*, vol. abs/1607.0, 2016.
28 H. Oosterhuis, S. Ravi, and M. Bendersky, "Semantic Video Trailers," *CoRR - ICML 2016 Workshop on Multi-View Representation Learning*, vol. abs/1609.0, 2016.
29 I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.
30 J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009)*, (Miami Beach, FL.), 2009.
31 K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," tech. rep., 9 2014.
32 K. Zhang, W.-l. Chao, F. Sha, and K. Grauman, "Supplementary Material : Video Summarization with Long Short-term Memory," *Eccv*, vol. abs/1605.0, pp. 1–7, 2016.
33 F. Chollet, "Xception: Deep Learning with Separable Convolutions," *arXiv preprint arXiv:1610.02357*, pp. 1–14, 2016.
34 C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," *CoRR*, vol. abs/1512.0, 2015.
35 K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *CoRR*, vol. abs/1512.0, 2015.
36 G. Huang, Z. Liu, and K. Q. Weinberger, "Densely Connected Convolutional Networks," *CoRR*, vol. abs/1608.0, 2016.

6 Appendices

Table 6 F-measures for each DCN model using combined diversity ϑ (higher is better). For each video in SumMe dataset, we show best result (bold). Finally we show the mean and standard deviation of the f-measures obtained by each DCN model.

		Computational Method ($\vartheta_i = \frac{1}{2}(D_i + K_i)$): Evaluation of DCN models						
Category	**Videoname**	**VGG16[31]**	**VGG19[31]**	**Xception [33]**	**InceptionV3 [34]**	**ResNet50 [35]**	**InceptionResNetV2**	**DenseNet [36]**
Egocentric	Base jumping	0,182	0,114	0,187	**0,200**	0,175	0,174	0,166
	Scuba	0,142	**0,230**	0,300	0,112	0,126	0,170	0,182
	Bike Polo	0,100	0,193	0,225	**0,290**	0,076	0,153	0,219
	Valparaiso_Downhill	**0,306**	0,283	0,216	0,260	0,235	0,203	0,233
Moving	Bearpark_climbing	0,088	0,107	0,173	0,195	0,133	**0,229**	0,147
	Bus_in_Rock_Tunnel	0,081	0,083	0,109	0,101	0,089	0,091	**0,117**
	Car_railcrossing	**0,117**	0,130	0,037	0,123	0,066	0,080	0,058
	Cockpit_Landing	0,140	0,139	0,124	0,126	0,189	**0,190**	0,147
	Cooking	0,075	0,128	0,132	0,204	**0,266**	0,207	0,265
	Eiffel Tower	**0,219**	0,152	0,147	0,101	0,144	0,135	0,129
	Excavators river crossing	0,092	0,118	0,086	0,081	0,056	**0,098**	0,089
	Jumps	0,063	0,049	0,051	0,175	0,040	0,044	**0,387**
	Kids_playing_in_leaves	0,339	0,263	**0,415**	0,319	0,390	0,221	0,196
	Playing_on_water_slide	0,040	0,046	0,048	0,074	0,055	**0,115**	0,047
	Saving dolphines	0,116	0,113	0,066	0,114	0,126	0,120	**0,165**
	St Maarten Landing	0,581	0,563	0,557	0,469	**0,610**	0,504	0,552
	Statue of Liberty	0,082	0,103	0,116	0,114	0,093	**0,152**	0,123
	Uncut_Evening_Flight	0,216	0,178	0,269	**0,300**	0,116	0,248	0,153
	paluma_jump	0,114	0,100	0,104	**0,259**	0,255	0,120	0,106
	playing_ball	0,142	0,092	0,152	0,097	0,053	**0,190**	0,154
	Notre_Dame	0,105	0,103	0,128	0,137	**0,144**	0,113	0,139
Static	Air_Force_One	0,375	**0,385**	0,263	0,348	0,356	0,145	0,222
	Fire Domino	0,094	**0,231**	0,155	0,103	0,215	0,121	0,099
	car_over_camera	0,426	0,414	**0,436**	**0,436**	0,435	0,413	0,414
	Paintball	**0,485**	0,471	0,478	0,480	0,378	0,423	0,432
mean score		0,189	0,192	0,199	**0,209**	0,193	0,186	0,198

Table 7 F-measures for different computational methods (higher is better). For each video in SumMe dataset, best results are shown in bold. Finally we show the mean and standard deviation of the f-measures obtained by each computational method.

		Human Annotators			**Computational Method**			
Category	**Videoname**	**Min**	**Mean**	**Max**	**Random**	**Gygli (2014)**	**Otani (2016)**	**Ours (DCN: Inception V3)**
Egocentric	Base Jumping	0,113	0,257	0,396	0,144	0,121	0,077	**0,200**
	Bike Polo	0,190	0,322	0,436	0,134	**0,356**	0,235	0,290
	Scuba	0,109	0,217	0,302	0,138	**0,184**	0,154	0,112
	Valparaiso Downhill	0,148	0,272	0,400	0,142	0,242	0,258	**0,260**
Moving	Bearpark climbing	0,129	0,208	0,267	0,147	0,118	0,178	**0,195**
	Bus in Rock Tunnel	0,126	0,198	0,270	0,135	0,135	**0,151**	0,101
	Car Rail Crossing	0,245	0,357	0,454	0,140	**0,362**	0,328	0,123
	Cockpit Landing	0,110	0,279	0,366	0,136	**0,172**	0,165	0,126
	Cooking	0,273	0,379	0,496	0,145	0,321	**0,329**	0,204
	Eiffel Tower	0,233	0,312	0,426	0,130	**0,295**	0,174	0,101
	Excavators River Crossing	0,108	0,303	0,397	0,144	**0,189**	0,134	0,081
	Jumps	0,214	0,483	0,569	0,149	**0,427**	0,015	0,175
	Kids Playing in Leaves	0,141	0,289	0,416	0,139	0,089	0,278	**0,319**
	Playing on Water Slide	0,139	0,195	0,284	0,134	**0,200**	0,183	0,074
	Saving dolphines	0,095	0,188	0,242	0,144	**0,145**	0,121	0,114
	St Maarten Landing	0,365	0,496	0,606	0,143	0,313	0,015	**0,469**
	Statue of Liberty	0,096	0,184	0,280	0,122	**0,192**	0,143	0,114
	Uncut Evening Flight	0,206	0,350	0,421	0,131	0,271	0,168	**0,300**
	Paluma Jump	0,346	0,509	0,642	0,139	0,181	**0,428**	0,259
	Playing Ball	0,190	0,271	0,364	0,145	0,174	**0,194**	0,097
	Notre Dame	0,179	0,231	0,287	0,137	**0,235**	0,093	0,137
Static	Air Force One	0,185	0,332	0,457	0,144	0,318	0,316	**0,348**
	Fire Domino	0,170	0,394	0,517	0,145	**0,130**	0,022	0,103
	Car Over Camera	0,214	0,346	0,418	0,134	0,372	0,132	**0,436**
	Paintball	0,145	0,399	0,503	0,127	0,320	0,274	**0,480**
Mean		0,179	0,311	0,409	0,139	**0,234**	0,183	0,209
Relative to human avg.		58%	100%	131%	45%	**75%**	59%	67%
Relative to human max.		44%	76%	100%	34%	**57%**	45%	51%

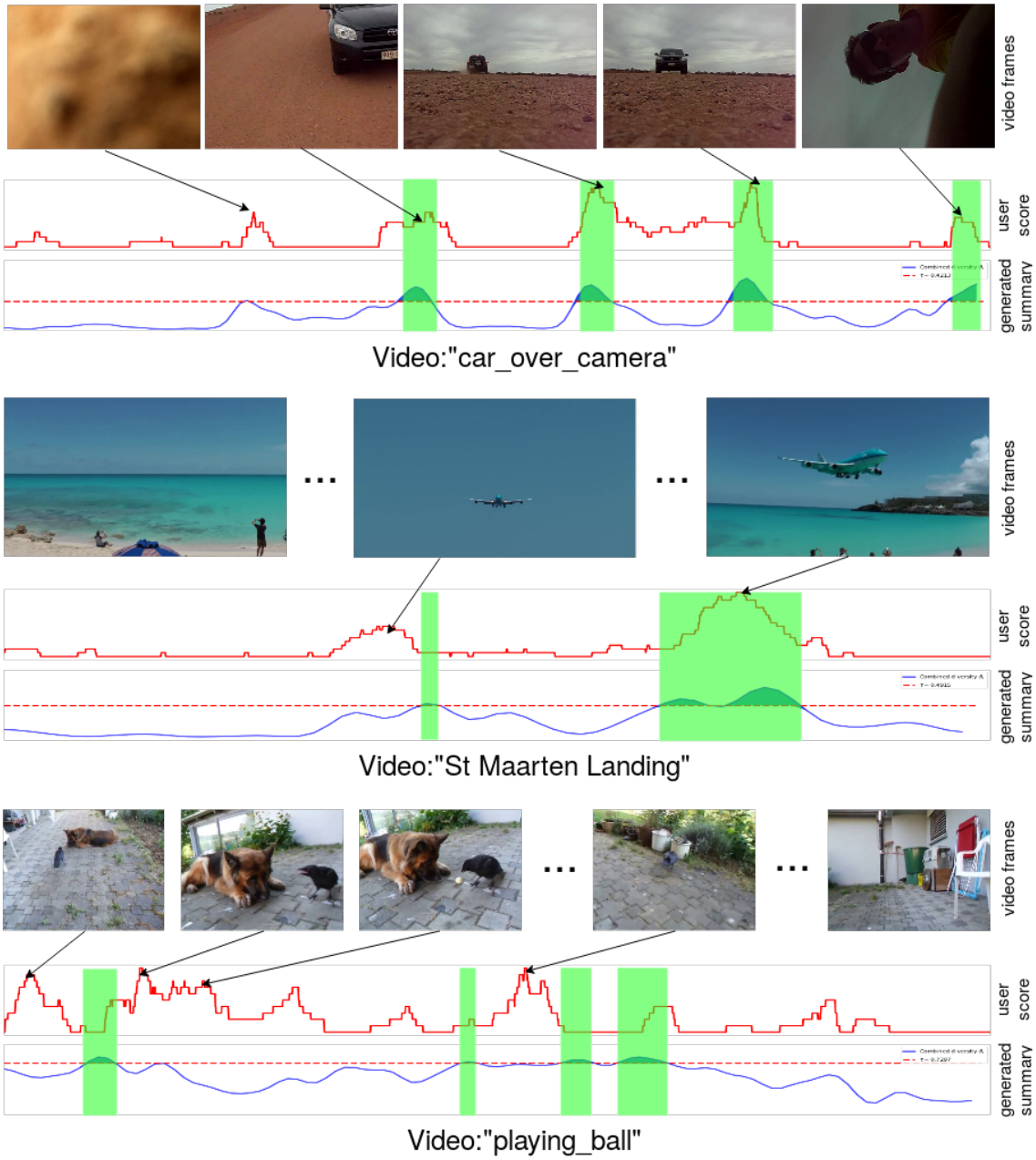

Fig. 7: **Example summaries**. Three video examples from dataset SumMe. For each video it is presented the mean user score in red, generated summary (at 15%) by our method in blue, and intersection of generated summary and user scores in green. It is also shown, video frames with high importance to human annotators.