MINIMUM-RISK STRUCTURED LEARNING OF VIDEO SUMMARIZATION

Fairouz Hussein, Massimo Piccardi

Faculty of Engineering and IT, University of Technology Sydney, Australia

{Fairouz.Hussein@student.,Massimo.Piccardi@}uts.edu.au

ABSTRACT

Video summarization is an important multimedia task for applications such as video indexing and retrieval, video surveillance, human-computer interaction and video "story-boarding". In this paper, we present a new approach for automatic summarization of video collections that leverages a structured minimum-risk classifier and efficient submodular inference. To test the accuracy of the predicted summaries we utilize a recently-proposed measure (V-JAUNE) that considers both the content and frame order of the original video. Qualitative and quantitative tests over two action video datasets - the ACE and the MSR DailyActivity3D datasets - show that the proposed approach delivers more accurate summaries than the compared minimum-risk and syntactic approaches.

Index Terms— Video summarization, minimum-risk classifiers, submodular functions, loss functions, structural SVM, V-JAUNE.

I. INTRODUCTION AND RELATED WORK

The amount of publicly-available video footage is growing at unprecedented rates thanks to the commoditization of video acquisition and the role played by social media. According to VIDCON 2015[1], YouTube users upload more than 400 hours of video to the site every minute. Moreover, SocialMediaToday has recenlty reported that video views on Facebook are averaging 8 billion a day[2]. With the rapidly expanding size of video repositories, the need for summarization tools is becoming more urgent. Fortunately, typical video content can be effectively summarized to a remarkable extent. For example, in sports videos an informative summary may contain highlights of scored points and defensive actions. In general, video summarization offers an efficient approach to abstract the main actions, scenes, or objects in a video to provide an easily-understood synopsis [1]. Over the years, a large number of algorithms have been proposed for automated summarization, aimed at both accuracy and efficiency. These algorithms can be mainly categorized as a) clustering approaches and b) frame-differences approaches. Overall, the main requirements of an effective video summary are well understood and reduce to adequate coverage of the original footage together with limited redundancy amongst the selected frames.

In this paper, we tackle the problem of video summarization by a minimum-risk structural approach. The advantages of minimum-risk approaches is that they train the predictor by minimizing loss functions which reflect the end-user expectation on performance. At the same time, the advantages of structural approaches is that they encapsulate the relations between frames rather the predicting each frame independently. As approach, we have chosen structural SVM for its established reputation as an accurate classifier [2]. With this approach, we provide efficient inference by designing efficient submodular functions. In addition, we contribute to the definition of an appropriate measure for the quality of the produced summaries. A popular metric for evaluating summaries for the more usual case of text data is a recall-based measure called ROUGE. Its rationale is to measure the similarity between the summary and the original text in terms of similarity of word histograms [3]. Inspired by ROUGE, [4] have introduced a metric called V-ROUGE for the evaluation of summaries of image collections. While this metric can also be used to measure the quality of a video summary, it does not take into account the order in which the frames are expected to appear. Therefore, in this paper we utilize a recently-proposed measure, V-JAUNE, which accounts for the similarity of both the frames' content and their order [5]. The experimental results over two challenging action video datasets, ACE [6] and MSR Daily-Activity3D [7], show that the proposed approach is capable of delivering accurate summaries both from a quantitative and qualitative perspective.

II. SUMMARIZATION VIA STRUCTURED LEARNING

In this section, we describe the summarization framework, including the submodular inference and the loss function. For a description of the structural SVM framework, the reader is referred to our recent paper [5].

[1] http://www.reelseo.com/vidcon-2015-strategic-insights-tactical-advice/

[2] http://www.socialmediatoday.com/marketing/top-5-facebook-video-statistics-2016-infographic

II-A. Submodular Inference

To formalize the problem, let us first note the sequence of frames of a video as $x = \{x_1, \dots x_i, \dots x_T\}$, and a corresponding sequence of binary variables indicating whether a frame is included in the summary or not as $y = \{y_1, \dots y_i, \dots y_T\}$. We also note a scoring function that assigns a "compatibility" score to x and y as $F(x, y)$ (the higher the score, the more appropriate is summary y for video x). Formally, we aim to infer an optimal summary, $\bar{y}$, of given size B (a "budget"):

$$\bar{y} = \underset{y}{\operatorname{argmax}} F(x, y) \quad s.t. \sum_{i=1}^{T} y_i = B \tag{1}$$

We now restrict the choice of scoring function to the case of linear models:

$$F(x, y) = w^\top \psi(x, y) \tag{2}$$

with w a parameter vector of non-negative elements and $\psi(x, y)$ a suitable feature function of equal size. The proposed feature function can be written as:

$$\psi(x, y) = \sum_{i,j=1, j \neq i}^{T} \lambda(y_i, y_j) s(x_i, x_j) \tag{3}$$

where:

$$\lambda(y_i, y_j) = \begin{cases} \lambda_1 > 0, & y_i = 1, y_j = 0 \text{ (coverage term)} \\ \lambda_2 < 0, & y_i = 1, y_j = 1 \text{ (non-redundancy term)} \\ 0, & y_i = 0, y_j = 0 \end{cases} \tag{4}$$

with $s(x_i, x_j)$ an arbitrary similarity function between frames x_i and x_j (for instance, the cosine similarity). In equation (4), the λ_1 terms represent the coverage terms since they reward the similarity between the summary frames and the remaining frames, while the λ_2 terms represent the non-redundancy terms since they penalize similar frames within the summary. From a computational perspective, the main advantage of a scoring function such as (3) is that it is submodular. Monotonic submodular functions enjoy important performance bounds for inference: a simple, greedy algorithm that picks the frames for the summary one at a time is guaranteed to achieve at least $((e-1)/e) \approx 0.632$ of the true maximum of the scoring function. Summarization functions have been proven to be both submodular and monotonic for reasonably small summaries, and the maximum returned by the greedy algorithm often exceeds the performance bound. However, they do not take into account the order in which the frames appear in the sequence. As a consequence, the summaries for actions such as "sitting down" and "standing up" may be indistinguishable. To ensure that the frames' sequentiality is instead properly taken into account, we propose to augment (3) as follows:

$$\psi(x, y) = [\sum_{i,j=1, j \neq i}^{T} \lambda(y_i, y_j) s(x_i, x_j) \mid \underbrace{\Omega(y)}_{\text{order term}}] \tag{5}$$

where:

$$\Omega(y) = \lambda_3 \sqrt{\sum_{i,j=1}^{T} (i-j)^2}, \quad y_i = 1, y_j = 0, \lambda_3 > 0 \tag{6}$$

In this way, a new term, $\Omega(y)$, is concatenated to the scoring function to reward the coverage of the frame indexes (notation $[a|b]$ represents the concatenation of a and b). This term helps ensure that the summary will contain a good representation of the frames based not only on their content, but also on their order in the sequence. The square root in (6) retains the submodularity of its argument, which is a conventional coverage term. This new term is a scalar, so the size of ψ (and thus w) only increases by one unit.

II-B. Learning with the V-JAUNE Loss

The linear model of (2-6) is learned with structural SVM [2]. For reasons of space, we refer the reader to [5] for details. The objective function for the training of structural SVM should use a loss function that properly reflects a desirable summarization. For this reason, we have adopted the recently-proposed V-JAUNE loss [5]. To describe it hereafter, we use a more compact notation, $y = \{y_1, \dots y_i, \dots y_B\}$, for a summary, consisting of the frame indexes of its B frames. Given a ground-truth summary, y^g, and a predicted summary, y, the loss function is defined as follows:

$$\begin{aligned} \Delta(y^g, y) &= \sum_{i=1}^{B} \delta(y_i^g, y_i) \\ \delta(y_i^g, y_i) &= \min\{\left\| x_{y_j^g} - x_{y_i} \right\|^2\}, \quad s.t. \quad i - \ell \leq j \leq i + \ell \end{aligned} \tag{7}$$

With this definition, loss function $\Delta(y^g, y)$ reflects the sequential order of the frames in the ground-truth and predicted summaries, while allowing for a $\pm\ell$ tolerance in the matching of the corresponding positions. This loss can be easily extended to account for multiple ground truths [5]. In addition, it is also advantageous from a computational perspective since it ensures that the key maximization of structural SVM (the so-called "loss-augmented inference") is submodular and therefore efficient. A concise proof is given hereafter.

Proposition: Function $w^\top \psi(x^n, y) + \Delta(y^n, y)$ needed by structural SVM for the loss-augmented inference is submodular.

Proof: Given two summaries, y_1 and y_2 with $y_1 \subset y_2$, and a new element, v, a function F is called submodular

if $F(y_1 \cup v) - F(y_1) \geq F(y_2 \cup v) - F(y_2)$. Function $\Delta(y^g, y)$ is submodular since it satisfies this constraint using the equal sign. At its turn, function $w^\top \psi(x^n, y)$ was proven to be submodular in [8]. Given that the sum of submodular functions is also submodular [9], the proposition follows. □

III. EXPERIMENTAL RESULTS

To evaluate the effectiveness of the proposed method, we have performed experiments on two challenging action datasets of depth videos: the Actions for Cooking Eggs (ACE) dataset [6] and the MSR DailyActivity3D dataset [7]. For both datasets, we have used comparable implementation settings: for each video, we have extracted dense local descriptors (HOG/HOF) over a regular spatio-temporal grid using the code from [10] resulting in 162-D individual descriptors. As feature encoding, we have used VLAD [11] which embeds the distance between the pooled local descriptors and the centres of a set of clusters. To obtain the encoding, we have first run k-means clustering over all the descriptors in the training set, empirically choosing $k = 64$ for ACE (clipped version) and $k = 32$ for ACE (unclipped version) and MSRDailyActivity3D. Then, for each frame, we have used the found clusters to encode all the frame's descriptors in an encoding of $162 \times k$-D dimensions, to be used as the measurement vector for the frame. As approaches, we have compared 1) the proposed system with a scoring function that uses only the content coverage and the non-redundancy terms (i.e., a "set" scoring function with $\lambda_3 = 0$ as in Lin and Bilmes [8]); 2) the proposed system with the full scoring function; and 3) the sum of absolute differences (SAD), a popular summarization approach which has been widely used in object recognition and video compression [12]. As software for the structural SVM model, we have used Joachims' solver [2]. As parameters, we have used summary size $B = 10$, regularization coefficient $C = 100$, and performed a grid search over the training set for weights $\lambda_1, \lambda_2, \lambda_3$ in absolute range $[0, 1]$ in 0.5 steps.

III-A. ACE

This dataset was collected in a simulated kitchen scenario using a Kinect camera at 30 fps and 640×480 resolution. The videos portray five actors cooking eggs according to five recipes: ham and eggs, scrambled eggs, boiled eggs, omelet and Kinshi-Tamago (a Japanese egg crepe). There are 35 videos in total, each ranging between $2,000$ and $12,000$ frames. The cooking entails eight different classes of actions: cutting, seasoning, peeling, boiling, turning, baking, mixing and breaking, annotated in the videos at frame level. For this dataset, we have performed summarization both at the video level ("unclipped" version) and at the action level ("clipped version). For the latter, we have clipped the individual cooking actions, obtaining 256 instances, each ranging between 20 and $4,469$ frames. For both versions, we have adopted the same training and test split that were proposed by the dataset's authors in [6]. In addition, we have asked five annotators (three for the clipped instances and two for the unclipped) to independently select $B = 10$ frames from each video as their preferred summary.

Results (clipped version). To evaluate the summaries, we have applied a quantitative comparison using the V-JAUNE loss. Table I reports the results obtained from the compared methods and by training with different ground-truth annotations. These results seem encouraging since the proposed method has achieved a lower loss value (0.891) than both the original scoring function of [8] which does not take into account the frame order (0.911) and SAD (0.927).

Table I. The V-JAUNE loss for the ACE dataset (clipped version).

Method	gt_1	gt_2	gt_3
SAD	0.927		
Lin and Bilmes [8] ($\lambda_3 = 0$)	0.911	0.919	0.921
Proposed method	0.894	0.906	**0.891**

Results (unclipped version). Table II shows the loss results from the compared methods. The best result (1.075) has been obtained, again, with the proposed method, with [8] and SAD ranking second and third, respectively. For a qualitative comparison, Figure 1 displays the summaries for recipes "ham and egg" and "omelet" obtained with the proposed method and SAD: in our judgment, the summaries provided by the proposed approach seem to better describe the entire preparation of the recipe. For example, frames from actions "seasoning" and "mixing" only appear in the summaries provided by the proposed method, and in the expected order.

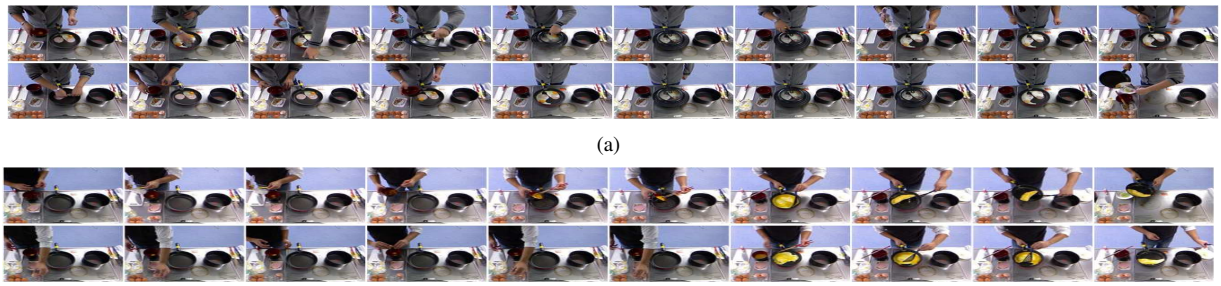
Table II. The V-JAUNE loss for the ACE dataset (unclipped version).

Method	gt_1	gt_2
SAD	1.098	
Lin and Bilmes [8] ($\lambda_3 = 0$)	1.088	1.096
Proposed method	**1.075**	1.088

III-B. MSR

The MSR DailyActivity3D dataset [7] is a Kinect dataset that depicts 16 common living-room activities (drinking, eating, reading and so forth). The total number of videos is 320, staged by 10 actors and performed in two different poses, one standing close to a couch and the other sitting on it. For evaluation, we have adopted the common cross-subject evaluation that uses subjects $1 - 5$ for training and subjects $6 - 10$ for testing.

Results. Table III reports the values of a denormalized V-JAUNE measure for the compared approaches (for this



(a)

(b)

Fig. 1. Examples of predicted summaries from the ACE dataset (unclipped version) for recipes a) *ham and egg* and b) *omelet*. In each subfigure, the first row is from the proposed method, the second from SAD.

dataset, we only have one annotation and cannot perform the usual multi-annotator V-JAUNE normalization [5]). Once more, the loss with the proposed method has been the lowest, followed by [8] and SAD.

Table III. The V-JAUNE loss for the MSR DailyActivity3D dataset.

Method	V-JAUNE (denorm)
SAD	5.652
Lin and Bilmes [8] ($\lambda_3 = 0$)	5.615
Proposed method	**5.497**

IV. CONCLUSION

In this paper, we have presented a novel approach for automated video summarization that leverages submodular inference and structural SVM. The two main contributions of the proposed approach have been: a) a submodular scoring function that appropriately takes into account the sequentiality of the frames, and b) the use of a dedicated, relevant loss (V-JAUNE [5]) for the training of structural SVM. The experimental results over two contemporary action datasets, ACE and MSR DailyActivity3D, have shown that the proposed approach has led to summaries of higher quality than those provided by an existing scoring function [8] and the sum of absolute differences (SAD) [12] in all cases.

V. ACKNOWLEDGMENTS

The authors wish to thank the Hashemite University for its generous financial support to this research.

VI. REFERENCES

[1] Y. Cong, J. Yuan, and J. Luo, "Towards scalable summarization of consumer videos via sparse dictionary selection," *IEEE Transactions on Multimedia*, vol. 14, no. 1, pp. 66–75, 2012.

[2] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun, "Large margin methods for structured and interdependent output variables," in *Journal of Machine Learning Research*, 2005, pp. 1453–1484.

[3] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out: Proceedings of the ACL-04 workshop*, vol. 8, 2004.

[4] S. Tschiatschek, R. K. Iyer, H. Wei, and J. A. Bilmes, "Learning mixtures of submodular functions for image collection summarization," in *NIPS*, 2014, pp. 1413–1421.

[5] F. Hussein and M. Piccardi, "V-JAUNE: A framework for joint action recognition and video summarization," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 13, no. 2, pp. 20:1–20:19, 2017.

[6] A. Shimada, K. Kondo, D. Deguchi, G. Morin, and H. Stern, "Kitchen scene context based gesture recognition: A contest in ICPR 2012," in *Advances in depth image analysis and applications*, 2013, pp. 168–185.

[7] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *CVPR*, 2012, pp. 1290–1297.

[8] H. Lin and J. Bilmes, "A class of submodular functions for document summarization," in *ACL*, 2011, pp. I:510–520.

[9] F. R. Bach, "Learning with submodular functions: A convex optimization perspective," *Foundations and Trends in Machine Learning*, vol. 6, no. 2-3, pp. 145–373, 2013.

[10] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," in *BMVC*, 2009, pp. 124.1–11.

[11] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *CVPR*, 2010, pp. 3304–3311.

[12] Z. Xiong, R. Radhakrishnan, A. Divakaran, Y. Rui, and T. S. Huang, *A unified framework for video summarization, browsing, and retrieval with applications to consumer and surveillance video*. Elsevier/Academic Press, 2006.

[13] A. Vedaldi, "A MATLAB wrapper of $\text{SVM}^{\text{struct}}$," http://www.vlfeat.org/~vedaldi/code/svm-struct-matlab.html, 2011.