JOINT ACTION RECOGNITION AND SUMMARIZATION BY SUB-MODULAR INFERENCE

Fairouz Hussein, Sari Awwad, Massimo Piccardi

Faculty of Engineering and IT, University of Technology Sydney, Australia
{Fairouz.Hussein@student.,Sari.Awwad@student.,Massimo.Piccardi@}uts.edu.au

ABSTRACT

Action recognition and video summarization are two important multimedia tasks that are useful for applications such as video indexing and retrieval, video surveillance, human-computer interaction and home intelligence. While many approaches exist in the literature for these two tasks, to date they have always been addressed separately. Instead, in this paper we move from the assumption that these two tasks should be tackled as a joint objective: on the one hand, action recognition can drive the selection of meaningful and informative summaries; on the other, recognizing actions from a summary rather than the entire video can in principle reduce noise and prove more accurate. To this aim, we propose a novel approach for joint action recognition-summarization based on the performing latent structural SVM framework, together with an efficient algorithm for inferring the action and the summary based on the property of sub-modularity. Experimental results on a challenging benchmark, MSR DailyActivity3D, show that the approach is capable of achieving remarkable action recognition accuracy while providing appealing video summaries.

Index Terms— Action recognition, video summarization, sub-modular functions, latent structural SVM, depth cameras.

1. INTRODUCTION AND RELATED WORK

Action recognition in video has been an important research area of multimedia signal processing for over a decade. Applications are varied and include, amongst others, video surveillance, human-computer interaction, sport analysis and home intelligence. Over the years, a variety of approaches have been proposed for recognition, including bag-of-features representations, sequential classifiers and deformable part models [1–4]. Such approaches have led to important results even in challenging cases with realistic scenarios and large class sets [5]. However, action recognition in video is still intrinsically challenged by the typical, extensive variations in illumination and view point. Fortunately, the recent release of inexpensive depth cameras such as Microsoft Kinect has helped mitigate these issues by adding an extra dimension to the traditional RGB components and generally improving recognition accuracy [6–8].

Another foundational area in multimedia signal processing is video summarization which provides concise information about a video by a few, informative frames. Video summaries can be used for indexing and retrieval or for storyboarding the videos to end users [9–11]. A useful video summary typically enjoys two properties: *coverage*, accounting for the similarity between the summary and the rest of the video, and *non-redundancy*, accounting for the diversity among the frames in the summary. These two properties can be combined into a single scoring function so as to assign a unique score to each candidate summary. Unfortunately, the number of possible candidates is exponential in the number of frames and an exhaustive search for the optimal summary is impossible. However, recent work from Lin and Bilmes [12], Sipos *et al*. [13], and Tschiatschek *et al*. [14] has remarked that the scoring function is sub-modular, and have exploited the properties of sub-modularity to provide fast and effective summary inference.

Given their intrinsic complexity, both action recognition and summarization can benefit from *structured prediction* approaches. Structured prediction leverages the formalism of graphical models to provide prediction for objects such as sequences, trees and graphs [15]. Its typical applications in image and video analysis range from image segmentation and action recognition to video indexing and summarization [2, 14, 16]. An increasingly popular approach in this area is structural SVM (SSVM) that is an extension of the conventional support vector machine for the classification of structured objects [17, 18]. SSVM has reported a strong experimental performance when compared to alternative approaches such as generative models and conditional random fields [15, 19].

To the best of our knowledge, action recognition and video summarization have been tackled to date as separate objectives. Instead, we believe that they could be usefully merged into a single, joint objective following the intuition that action recognition can drive the selection of meaningful frames for the summary and that, in turn, recognizing the action from a summary rather than the entire video may reduce noise and prove more accurate. Therefore, in this paper we present an approach based on latent structural SVM

that jointly provides the action class and the summary for an action video. Our main contribution is the design of a novel scoring function which enjoys the property of sub-modularity and therefore supports efficient inference of both the action and the summary. We present experiments over a challenging benchmark, MSR DailyActivity3D, showing that the approach is capable of achieving remarkable action recognition accuracy while providing meaningful and visually-appealing video summaries.

2. RECOGNITION AND SUMMARIZATION BY SUB-MODULAR FUNCTIONS

The goal of our work is to provide simultaneous classification and summarization of a video depicting an action. To this aim, let us note the sequence of measurements from the frames as $x = \{x^1,..,x^t,..,x^T\}$ where T is the sequence length; the sequence of binary variables indicating whether a frame belongs to the summary or not as $h = \{h^1,..,h^t,..,h^T\}$; and the action class as y. Formally, we aim to jointly infer class label y and summary h while keeping the summary within a given, maximum size, B:

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

Lin and Bilmes in [12] have shown that desirable summaries (i.e., summaries with good coverage and limited redundancy) enjoy the property of *sub-modularity*. Sub-modularity can be intuitively explained as a law of diminishing returns [12]: let us assume to have a scalar function, F, which can measure the quality of a given summary, together with an arbitrary summary, A. We now add a new element, v, to A and compute the difference in value between $F(A \cup v)$ and $F(A)$ (the "return" of v for A). Let us then consider a super-set of A, $B \supset A$, and add v to it: sub-modularity holds if the return of v for B is less than or equal to the return of v for A. In simple terms, the larger the summary is, the less is the benefit brought in by a new element. This property can be formally expressed as:

$$\forall A \subset B, v : F(A \cup v) - F(A) \geq F(B \cup v) - F(B) \tag{2}$$

Note that sub-modular functions are not required to be monotonically non-decreasing, i.e., returns can be negative; however, (2) must hold. For simplicity, in the following we also assume F to be non-negative. The attractive property of sub-modularity is that a value for F with a guaranteed lower bound can be found by simply selecting the elements for the summary one by one. The approximate maximum returned by such a greedy algorithm is guaranteed to be at least $(1 - 1/e) \approx 0.632$ of the actual maximum and is found to be often better in practice [12, 20].

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

$$F(x, h, y) = w^T \psi(x, h, y) \tag{3}$$

with w a parameter vector and $\psi(x, h, y)$ a suitable feature function of equal size. We further restrict w and $\psi(x, h, y)$ to be non-negative in all their elements. Lin and Bilmes in [12] have proposed the following feature function for summarization:

$$\psi(x, h, y) = \sum_{t=1}^{T} \left(\sum_{u=1}^{T} \delta(h^t, h^u) \sigma(x^t, x^u) \right) \tag{4}$$

where

$$\delta(h^t, h^u) = \begin{cases} \lambda_1 & \text{if } h^t = 1, h^u = 0 \\ -\lambda_2 & \text{if } h^t = 1, h^u = 1 \\ 0 & \text{otherwise,} \end{cases} \tag{5}$$

with $\lambda_1, \lambda_2 > 0$, and $\sigma(x^t, x^u)$ a non-negative function measuring the similarity between frames x^t and x^u. A frame x^t is selected for the summary if its corresponding binary indicator, h^t, is set to one. Therefore, the λ_1 terms in (4) are the coverage terms while the $-\lambda_2$ terms promote non-redundancy in the summary by penalising similar frames. Following [12], it is easy to prove that function (4) is sub-modular.

Functions based on between-frame similarities such as (4) are suitable for summarization, but do not properly describe the class of the action since their space is very sparse. Typical feature functions for action recognition are instead based on bagging or averages of the frame measurements. To provide joint summarization and recognition, we propose to modify (4) as follows:

$$\psi(x, h, y) = \sum_{t=1}^{T} \left(\delta(h^t) x^t + \sum_{u=1}^{T} \delta(h^t, h^u) \sigma(x^t, x^u) \right) \tag{6}$$

with

$$\delta(h^t) = \begin{cases} \lambda_3 > 0 & \text{if } h^t = 1 \\ 0 & \text{if } h^t = 0 \end{cases} \tag{7}$$

In this way, a new term is added containing the weighted sum of all measurements x^t in the summary. Such a term is equivalent to a pooled descriptor and promises to be informative for action recognition. We now prove that (6) is still sub-modular:

Proof: Given a current summary, h, the addition of any new frame to it makes term $\sum_{t=1}^{T} \delta(h^t) x^t$ vary by the same amount irrespectively of h. This term therefore satisfies inequality (2) with the equal sign. Given that convex combinations of sub-modular functions are also sub-modular [12], the overall sub-modularity of (6) follows. □

The main benefit of sub-modular scoring functions are the performance guarantees on greedy inference algorithms. Algorithm 1 shows the greedy algorithm that we use to jointly infer the best action class and the best summary, choosing one frame for the summary at a time.

Algorithm 1 Greedy algorithm for inferring class y^* and summary h^* given scoring function $F(x, h, y)$.

$max = -\infty, \text{argmax} = 0$
for $y = 1...|Y|$ **do**
  $h^* \leftarrow \emptyset$
  $X \leftarrow x$
  while $X \neq \emptyset$ **and** $|h^*| \leq B$ **do**
    $k \leftarrow \text{argmax}_{v \in X} F(x, h^* \cup v, y) - F(x, h^*, y)$
    $h^* \leftarrow h^* \cup \{k\}$
    $X \leftarrow X \backslash \{k\}$
  end while
  if $F(x, h^*, y) > max$ **then**
    $max = F(x, h^*, y)$
    $argmax = y$
  end if
end for

2.1. Latent Structural SVM

As framework for learning parameter vector w, we adopt the popular latent structural SVM [18] which has proved effective in a variety of multimedia signal processing applications [2, 16, 21]. In the training set, the action classes are supervised, but the summaries are completely unsupervised. Given a training set with N videos, (x_i, y_i), $i = 1 \dots N$, the learning objective of latent structural SVM:

$$\begin{aligned} w^* = & \underset{w, \xi_{1:N}}{\text{argmin}} \ \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{N} \xi_i \\ s.t. \ & w^T \psi(x_i, h_i^*, y_i) - w^T \psi(x_i, h, y) \geq \Delta(y_i, y) - \xi_i \\ & \forall \{y, h\} \neq \{y_i, h_i^*\} \end{aligned} \tag{8}$$

$$h_i^* = \underset{h}{\text{argmax}} \, w^{*T} \psi(x_i, h, y_i) \tag{9}$$

is an iterative objective that alternates between the constrained optimization in (8), performed using the current values for latent variables h_i^*, and a new assignment for h_i^* (9) from updated model w^*. The loss function that we choose to minimize, $\Delta(y_i, y)$, only accounts for the loss from action misclassifications. As such, the selection of frames for the summary, h, is solely driven by the requirement of maximizing the action recognition accuracy.

The optimization in (8) is a standard optimization that can be solved by use of any common solver. However, since the number of constraints in (8) is exponential, we adopt the relaxation of [17] which can find almost-correct solutions using only a polynomial-size working set of constraints. The working set is built by searching the sample's most violated constraint at each iteration of the solver:

$$\xi_i = \max_{y,h} (-w^T \psi(x_i, h_i^*, y_i) + w^T \psi(x_i, h, y) + \Delta(y_i, y)) \tag{10}$$

which equates to finding the labeling with the highest sum of score and loss:

$$\bar{y}_i, \bar{h}_i, = \underset{y,h}{\text{argmax}} (w^T \psi(x_i, h, y) + \Delta(y_i, y)) \tag{11}$$

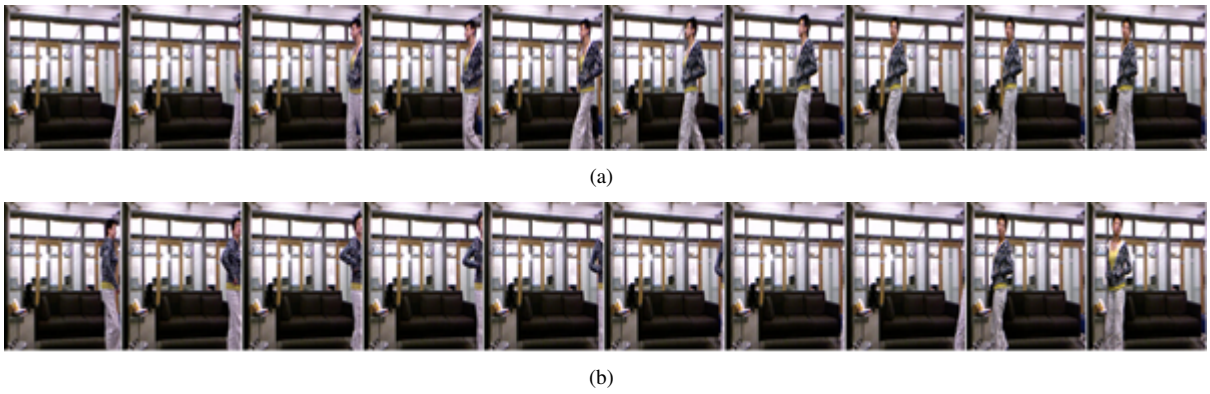
This problem is commonly referred to as "loss-augmented inference" due to its similarity to the standard inference and can be, again, solved by Algorithm 1 simply with the addition of loss $\Delta(y_i, y)$ to the score.

3. EXPERIMENTS

In this section, the proposed method is evaluated on the MSR DailyActivity3D dataset [6] released by Microsoft Research and captured using the Kinect RGBD camera. It depicts 16 common living-room activities including: drinking, eating, reading, using cell phones, writing, using computer/laptop, vacuuming, cheering up, sitting still, tossing crumbled paper, playing games, lying on the sofa, walking, playing the guitar, standing up, and sitting down. The total number of videos is 320, staged by 10 actors and performed in two different poses, one standing close to the couch and the other sitting on it. For evaluation, a cross-subject evaluation is common, with subjects $1 - 5$ used for training and subjects $6 - 10$ for test.

To pursue a more general approach, we have decided not to use the information about the actor's skeleton, limiting feature extraction to the depth and RGB streams. For each video, we have extracted local descriptors (HOG/HOF) over a regular spatio-temporal grid using the code from [1]. As time scale we have used $\tau = 2$ resulting in 162-D descriptors. For the encoding, we have first run k-means with $k = 32$ clusters from the entire set of descriptors of the training set. Then, we have encoded all the descriptors of each frame using VLAD [22] which embeds the distance between the pooled local features and the cluster's centres. The resulting encoding is a $162 \times 32 = 5,184$-D vector and is our measurement for the frame. As software for the latent structural SVM model, we have used Joachims' solver [17] with Vedaldi's MATLAB wrapper [23]. As parameters, we have used summary size $B = 10$, regularization coefficient $C = 100$ and performed a grid search over the training set for weights $\lambda_1, \lambda_2, \lambda_3$. [17, 23]

For performance evaluation, we care to note that our approach is the only approach to date to provide action recognition and video summarization as an integrated task. To evaluate the action recognition component, we compare the test-set recognition accuracy using depth videos with: 1) a reference system using the pooled descriptors from all frames



(a)

(b)

Fig. 1. Summary examples (displayed as RGB frames) for action *walk*: a) proposed method; b) SAD.

as measurement and libsvm as the classifier [24]; and 2) the proposed system using using the pooled descriptors from all frames as measurement, and without the summarization component (i.e., $\lambda_1 = \lambda_2 = 0$). In addition, we compare the action recognition accuracy with a system from the literature that uses dynamic time warping [25]; to the best of our knowledge, this is the only approach which does not use the actor's skeletal information in any form (locations or angles). Table 1 shows that the accuracy achieved with the proposed method (60.0%) is much higher than that of the reference system (34.4%) and also remarkably higher than that of the proposed method using all frames (48.8%). This proves that action recognition based on a selected summary can be more accurate than recognition from the entire video, and validates the intuition of providing action recognition and summarization jointly. In addition, the accuracy is also significantly higher than that from the dynamic time warping approach (54.0%). These accuracy levels can be regarded as satisfactory since they are far above chance accuracy, i.e. $1/16 = 6.25\%$ for this dataset. Eventually, Table 1 shows that the accuracy from depth videos is also remarkably higher than that from RGB videos (46.3%), showing that depth is a more informative clue for recognizing actions.

Table 1. Comparison of action recognition accuracy on the MSR Daily Activity 3D dataset.

Method	Accuracy
libsvm [24]	34.4%
Proposed method (all frames)	48.8%
Proposed method	**60.0**%
Dynamic temporal warping [25]	54.0%
Proposed method (RGB videos)	46.3%

For the evaluation of the summarization component, since a ground truth is not available, we resort to qualitative comparisons. In particular, we compare the summaries obtained with the proposed method with those produced by a popular summarization approach, the sum of absolute differences (SAD), which has been widely used in object recognition and video compression [26]. SAD is a low-level approach that selects the frames for the summary as those with the largest, absolute difference from the previous frame, up to the given budget. The examples displayed in Fig. 1 show that the summaries provided by the proposed approach appear more meaningful, faithful and informative about the content of the video.

4. CONCLUSION

In this paper, we have presented a joint approach for action recognition and summarization of action videos. The main benefit from the joint approach is two-fold: on the one hand, the video summaries are driven by high-level concepts such as actions and activities. On the other hand, the selection of relevant frames leads to improvements in action recognition accuracy. Experiments carried over a probing dataset (MSR DailyActivity3D) containing both RGB and depth videos have shown that:

- the accuracy achieved by the proposed approach on action recognition from depth videos is higher than that of alternative methods which do not deliver summaries as an objective or a by-product;
- in a qualitative comparison, the summaries achieved by the proposed approach appear to be more informative than those provided by a low-level approach such as SAD.

A future extension of this work could be the inclusion of partially-supervised summaries in the annotation, with the integration of summary loss functions such as V-ROUGE [14] in the learning objective.

5. REFERENCES

[1] Ivan Laptev, Marcin Marszalek, Cordelia Schmid, and Benjamin Rozenfeld, "Learning realistic human actions from movies," in *CVPR*, 2008.

[2] Minh Hoai, Zhen-Zhong Lan, and Fernando De la Torre, "Joint segmentation and classification of human actions in video," in *CVPR*, 2011, pp. 3265–3272.

[3] Kevin Tang, Li Fei-Fei, and Daphne Koller, "Learning latent temporal structure for complex event detection," in *CVPR*, 2012.

[4] Pedro F. Felzenszwalb, Ross B. Girshick, David A. McAllester, and Deva Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, 2010.

[5] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," *CoRR*, vol. abs/1212.0402, 2012.

[6] Jiang Wang, Zicheng Liu, Ying Wu, and Junsong Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *CVPR*, 2012, pp. 1290–1297.

[7] O. Oreifej and Zicheng Liu, "Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences," in *CVPR*, 2013, pp. 716–723.

[8] N.C. Tang, Yen-Yu Lin, Ju-Hsuan Hua, Ming-Fang Weng, and H.-Y.M. Liao, "Human action recognition using associated depth and skeleton information," in *ICASSP*, 2014, pp. 4608–4612.

[9] Yu-Fei Ma, Lie Lu, Hong-Jiang Zhang, and Mingjing Li, "A user attention model for video summarization," in *ACM Multimedia*, 2002, pp. 533–542.

[10] Yang Liu, Feng Zhou, Wei Liu, Fernando De la Torre, and Yan Liu, "Unsupervised summarization of rushes videos," in *ACM Multimedia*, 2010, pp. 751–754.

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

[12] Hui Lin and Jeff Bilmes, "A class of submodular functions for document summarization," in *ACL*. Association for Computational Linguistics, 2011, pp. I:510–520.

[13] Ruben Sipos, Pannaga Shivaswamy, and Thorsten Joachims, "Large-margin learning of submodular summarization models," in *EACL*, 2012, pp. 224–233.

[14] Sebastian Tschiatschek, Rishabh K Iyer, Haochen Wei, and Jeff A Bilmes, "Learning mixtures of submodular functions for image collection summarization," in *NIPS*, 2014, pp. 1413–1421.

[15] Sebastian Nowozin and Christoph H. Lampert, "Structured learning and prediction in computer vision," *Found. Trends. Comput. Graph. Vis.*, vol. 6, no. 3–4, pp. 185–365, Mar. 2011.

[16] Weilong Yang, Yang Wang, and Greg Mori, "Recognizing human actions from still images with latent poses," in *CVPR*, 2010, pp. 2030–2037.

[17] Ioannis Tsochantaridis, Thorsten Joachims, Thomas Hofmann, and Yasemin Altun, "Large margin methods for structured and interdependent output variables," in *Journal of Machine Learning Research*, 2005, pp. 1453–1484.

[18] Chun-Nam John Yu and Thorsten Joachims, "Learning structural svms with latent variables," in *ICML*, 2009, pp. 1169–1176.

[19] Yang Wang and Greg Mori, "Hidden part models for human action recognition: Probabilistic versus max margin," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 7, pp. 1310–1323, 2011.

[20] G.L. Nemhauser, L.A. Wolsey, and M.L. Fisher, "An analysis of approximations for maximizing submodular set functions-i," *Mathematical Programming*, vol. 14, no. 1, pp. 265–294, 1978.

[21] Yang Wang and Greg Mori, "Max-margin hidden conditional random fields for human action recognition," in *CVPR*, 2009, pp. 872–879.

[22] Hervé Jégou, Matthijs Douze, Cordelia Schmid, and Patrick Pérez, "Aggregating local descriptors into a compact image representation," in *CVPR*, 2010, pp. 3304–3311.

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

[24] Chih-Chung Chang and Chih-Jen Lin, "Libsvm: a library for support vector machines," *ACM Trans. Intell. Syst. Tech.*, vol. 2, no. 3, pp. 27, 2011.

[25] Meinard Müller and Tido Röder, "Motion templates for automatic classification and retrieval of motion capture data," in *ACM/SIGGRAPH SCA*, 2006, pp. 137–146.

[26] Ziyou Xiong, Regunathan Radhakrishnan, Ajay Divakaran, Yong Rui, and Thomas S. Huang, *A unified framework for video summarization, browsing, and retrieval with applications to consumer and surveillance video*, Elsevier/Academic Press, 2006.