HUMAN ACTION RECOGNITION-BASED VIDEO SUMMARIZATION FOR RGB-D PERSONAL SPORTS VIDEO

Antonio Tejero-de-Pablos, Yuta Nakashima, Tomokazu Sato, Naokazu Yokoya

Nara Institute of Science and Technology, Japan
{antonio.tejero.ao4,n-yuta,tomoka-s,yokoya}@is.naist.jp

ABSTRACT

Automatic sports video summarization poses the challenge of acquiring semantics of the original video, and existing work leverages various knowledge in application domains, *e.g.*, structure of games and editing conventions. In this paper, we propose a personal sports video summarization method for self-recorded RGB-D videos, which became available to the public due to the commodification of off-the-shelf RGB-D sensors. We focus on sports whose games consist of a succession of actions and, unlike previous research, we use human action recognition on the depth sequences in order to acquire higher level semantics of the video. The recognition results are used along with an entropy-based activity measure to train a hidden Markov model of the highlights of different games to extract a summary from the original RGB-D video. We trained our novel highlights model with the subjective opinion of users with different experience in the sport. We took Kendo, a martial art, as an example sport to evaluate our method, and objectively/subjectively investigated the accuracy and quality of the generated summaries.

Index Terms— Video summarization, personal sports video, highlight extraction, RGB-D video, human action recognition

1. INTRODUCTION

Nowadays, a vast amount of personal videos are taken and stored due to the exponential growth of commercial devices capable of video recording. One of the main targets of these videos is sports that users may record in, *e.g.*, a public event, a professional game, or even their own performances. However, in many cases, these videos are just stored and never reviewed, partly because they are usually long, containing redundant and uninteresting parts.

Video summarization is a technique to compact the lengthy original videos for quick review [1]. There are approaches specialized in sports video, leveraging various types of knowledge on the target sport in order to facilitate video summarization. For example, broadcast programs are recorded and edited by an expert following editing conventions, such as standard camera viewpoints, narration, or superimposed text [2]. These editing conventions, which can be easily detected, are associated with higher level semantics and help to find the relevant parts of the video. Some sports like baseball and American football have a certain structure in a game itself [3, 4]. However, personal videos usually lack any kind of editing conventions and the structure of the sport is not always well-defined.

In this paper, we propose a method for personal sports video summarization using a new source of semantics extraction, *i.e.*, depth of scenes, which becomes available and affordable due to the recent development of RGB-D sensors including Microsoft Kinect. More specifically, some sports, such as tennis, boxing, and martial arts, consist of a series of actions (*e.g.*, uppercut, and jump-kick), and our method automatically labels them by applying human action recognition (HAR) to RGB-D video sequences. While HAR has been traditionally applied to color images [5], the use of depth video highly improves HAR accuracy and robustness against illumination changes, camera blurring, etc. [6]. We model the highlights of a game based on HAR results to extract them from a lengthy original RGB-D video.

The contributions of this work are summarized as follows:

- We propose a novel method for summarizing personal sports video based on HAR from a self-recorded RGB-D video sequence. To the best of our knowledge, this is the first attempt to use this kind of analysis for video summarization. Our method is suitable for sports that can be recorded at a close distance.
- We evaluate the performance of our method both objectively and subjectively to show its effectiveness and accuracy. We carried out a survey of users with and without experience in the sport to investigate the adequacy of our method to their particular preferences.

2. RELATED WORK

One of the major approaches to analyze sports video for summarization is to use editing conventions for broadcast programs, which are common to almost all videos of a specific sport [2]. In [7], the authors proposed an automatic framework for soccer video summarization based on editing conventions as well as detection of soccer field elements (*e.g.*, goal). Other works use these conventions to extract higher semantics [8, 9] and find highlights, which are video segments containing important events of the game [10]. In [3], important events in a certain class of sports (American football, baseball, and sumo wrestling) are modeled by "plays", defined according to the rules of the sports, which can be detected based on the conventional patterns in broadcast programs. Another way of extracting the semantic concepts in sports video is to use the metadata of the video content. In [4], Nitta *et al.* used the play information contained in the metadata of MPEG-7 (*i.e.*, the inning structure in baseball games). However, most of these approaches are domain dependent, which makes them hard to generalize to other sports [10].

Personal videos usually do not follow any editing convention, storyline, or image quality standards [11], and therefore the aforementioned methods are not suitable for them [12]. Clustering [13] is one of the most common approaches to summarize videos that can be applied to personal videos, but it is basically dedicated to reduce the redundancy of the video and may not consider any semantics. There are different approaches to sample interesting shots in a video, for which semantics may not be available, by calculating the activity level in its color frames that represents how lively the scene changes [11, 14]. This technique has been also applied to sports video to segment semantically relevant events in broadcast games of basketball, soccer, and tennis [15].

What is to be included in a summary is sometimes not obvious even if higher-level semantics are extracted. Some works model the video highlights based on the viewers' preferences, which can be obtained explicitly from viewers or inferred from their reactions while watching videos [16].

Similarly to the approaches that extract "plays" from edited videos, for some types of sports like Kendo, we can define a set of actions of individual players that make up the course of a game. For such videos, instead of using the particular structure of a sport or editing conventions, we can recognize players' actions directly and use them as higher level semantics for video summarization. Zhu *et al.* [17] attempted to use player's action recognition as a complement to editing conventions to acquire semantics. However, due to the difficulty in recognizing actions from RGB video frames, they adopted only two action classes. To improve the recognition performance, considering the recent commodification of sensors, we use RGB-D videos, which facilitate player segmentation, and reduce the impact of appearance variations and ambiguity in their actions [18].

3. HAR-BASED SPORTS VIDEO SUMMARIZATION

Figure 1 depicts an overview of our method, which takes an RGB-D sports video sequence and generates a summary containing the highlights of the game. The sequence is firstly segmented into T uniform-length (*i.e.*, 3 seconds) sub-sequences. In order to exploit the inherent semantics of the video, we apply HAR to each sub-sequence. In most sports, multiple players are involved in the game; therefore, HAR is also applied to each player to calculate the dissimilarity between the action of that player in each sub-sequence and each action instance in a predefined set of action classes. We use this dissimilarity and an activity measure, which quantifies the amount of motion in the sub-sequence, to model interesting sub-sequences that are to be included in the resulting highlights summary with a hidden Markov model with Gaussian mixture model emissions (GMM-HMM), which is trained with labeled sub-sequences. Finally the summary is extracted via skimming curve formulation [1] for a given time length L.

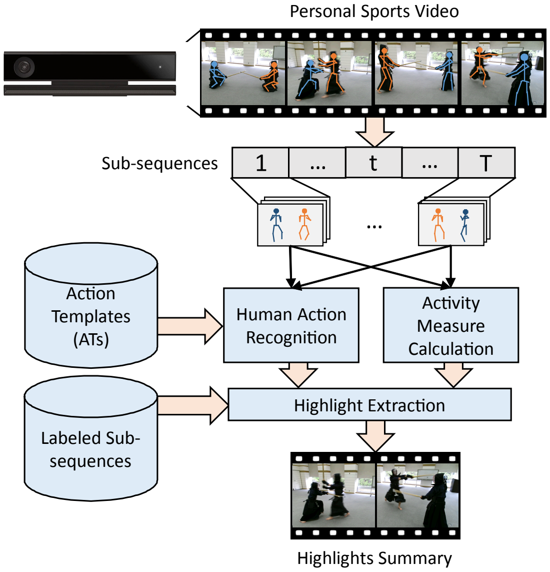


Fig. 1. Overview of our summarization method.

3.1. HAR via Action Templates

In order to calculate the dissimilarity between the action of players in the t-th sub-sequence and each of the predefined actions, we apply HAR to each player p. From the depth maps in a sub-sequence, we obtain the skeleton (*i.e.*, a set of 3D joint positions) of each player using a skeleton tracker ([19], for example) to gain robustness to view variations with respect to both the camera locations and subject appearances. We use a simple method for HAR [20], which calculates the distance between the sequence of skeletons of player p in a sub-sequence and each of the action templates (referred to as ATs) in an action dataset.

An AT is a set of action instances (sequences of skeletons) of a predefined action class specialized for the sport. To generate an AT, we extract the skeleton from a depth map sequence that contains one of the predefined actions. Skeleton trackers can also provide a confidence value for each estimated joint position. These positions are transformed to the player's coordinate system, whose origin is at one of the joints (*e.g.*, torso). The sequence of transformed skeletons along with the confidence values form the AT.

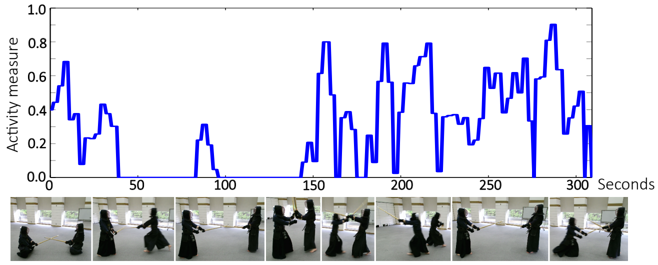


Fig. 2. Activity measure along the course of a Kendo game.

For the given t-th input sub-sequence, which may contain multiple players in unknown action classes, we apply a similar process to extract the players' skeletons and transform them into each player's coordinate system. We then calculate the distance between the sequence of skeletons for each player and each of the ATs. Since the duration of an action varies from instance to instance, we adopt dynamic time warping [21] to handle this. In this method, the confidence values are used to filter the noisy sections of the trajectories. Let N denote the number of the predefined actions classes and M the number of action instances per action class. Our HAR method generates a vector $\mathbf{d}_{tp}$ whose n-th element d_{tp}^{n} is given by $d_{tp}^{n} = \min_m d_{tp}^{nm}$, where d_{tp}^{nm} is the distance between player p's action in t-th sub-sequence and the m-th AT for the n-th action class ($m = 1, \ldots, M$ and $n = 1, \ldots, N$).

3.2. Activity measure

The HAR outputs may not reflect how sudden or prominent the actions are. In [15], they hypothesize that interesting highlights in sports video are characterized by certain patterns in the entropy of the intensities in RGB frames. For each subsequence, we use the activity measure of each player's motion based on the entropy of the motion of each joint. For this, we divide the 3D space of the player's coordinate system into V volumes and calculate the ratio r_v of the number of frames in the subsequence in which the joint j of player p fall into volume v. The entropy for joint j is given by

$$e_j = -\sum_{v=1}^{V} r_v \log(r_v). \tag{1}$$

We define the activity measure of a player as $a = \sum_{j=1}^{J} e_j$ where J is the total number of joints. Figure 2 shows the variation of a along time. The activity measure rises as sudden actions are executed successively, and decreases with repetitive motion (or lack of motion). Sections with zero activity are those where players were not recognized.

For sub-sequence t, we define a feature vector $\mathbf{f}_t^{\top} = (\mathbf{d}_{t1}^{\top}, a_{t1}, \mathbf{d}_{t2}^{\top}, a_{t2}, \ldots, \mathbf{d}_{tP}^{\top}, a_{tP})$, which is a concatenation of the HAR result $\mathbf{d}_{tp}$ and activity measure a_{tp} for all players, where P is the number of the players in the t-th sub-sequence and a_{tp} is the activity measure for player p.

3.3. Highlight extraction

In order to create the summary from the original sequence, we calculate the probability of each sub-sequence of being interesting/non-interesting based on the features, assuming that the segments that are labeled as interesting by users are the highlights of the game. We adopt a GMM-HMM to model interesting/non-interesting segments because adjacent sub-sequences are expected to be highly correlated.

In our method, we assume that the emission probability $\Pr(\mathbf{f}_t|e)$ of $\mathbf{f}_t$ given e follows a Gaussian mixture model, where $e = 1$ indicates that the sub-sequence belongs to an interesting segment and $e = 0$ otherwise. Specifically, the emission probability is given by

$$\Pr(\mathbf{f}_t|e) = \sum_{k=1}^{K} w_{ek} \mathcal{N}(\mathbf{f}_t|\mu_{ek}, \mathbf{\Sigma}_{ek}), \tag{2}$$

where w_{ek}, μ_{ek}, and Σ_{ek} are the mixture weight, the mean, and the covariance matrix of the k-th mixture component for state e. Letting $F = \{\mathbf{f}_t | t = 1, \ldots, T\}$ and $\mathbf{e}^{\top} = (e_1, \ldots, e_T)$, the probability $\Pr(F_T, \mathbf{e})$ is given by

$$\Pr(F, \mathbf{e}) = \Pr(e_0) \prod_{t=1}^{T} \Pr(e_t|e_{t-1}) \prod_{t=1}^{T} \Pr(\mathbf{f}_t|\mathbf{e}_t, \phi), \tag{3}$$

where $\Pr(e_0)$ is the initial state probability. We can calculate the posterior probability $\Pr(e_t|F)$ using the forward-backward algorithm. Since we have labeled videos for training, the parameters for initial state probability $\Pr(e_1)$ and the transition probability $\Pr(e_t|e_{t-1})$ can be easily determined by counting, and the parameters for GMM (*i.e.*, w_{ek}, μ_{ek}, and Σ_{ek}) can be estimated using the EM algorithm [22].

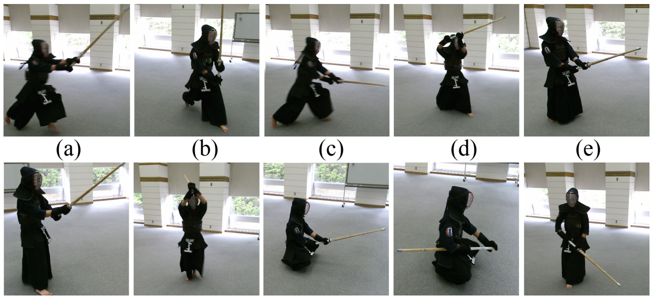
Once the probabilities are obtained, we generate the summary using skimming curve formulation [1]. Given a certain summary length L in seconds, we apply thresholding to $\Pr(e_t|F)$ by reducing the threshold until we find a set of segments whose total length in seconds is the largest below L. We arrange the extracted segments in temporal order to generate a video summary.

4. EXPERIMENTAL RESULTS

To evaluate our method, we chose Kendo as an example sport, which is a martial art featuring two players and a set of recognizable actions. Using a Microsoft Kinect v2 sensor, we recorded 10 RGB-D videos (90 minutes in total), which contain 12 combats. The videos used in the experiments were taken close to the players (2m–4m) for depth map acquisition. We used [19] for skeleton tracking. Apart from these videos, we generated a dataset for HAR, which contains 200 action instances (10 action classes×4 actors×5 repetitions) of action classes (a) *men*, (b) *kote*, (c) *dou*, (d) *bougyo*, (e) *kamae*, (f) *tsubazeriai*, (g) *hikimen*, (h) *sonkyo*, (i) *osametou*, and (j)

Table 1. Confusion matrix of [20] over the kendo dataset (%).

		Recognition results									
		(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)	(j)
Action classes	(a)	25	20	15	5	25			10		
	(b)		20	30	20	5	10				15
	(c)	15	10	50	5	10			5		5
	(d)	10		5	15	45	25				
	(e)	20			20	40	20				
	(f)	10			35	35	20				
	(g)	20		5				50		25	
	(h)	60	10						30		
	(i)	35				5		10	5	45	
	(j)	50	20	5						10	15


(a) (b) (c) (d) (e)

(f) (g) (h) (i) (j)

Fig. 3. Actions used in the dataset.

aruki. These actions consist of strikes in different body parts and defense positions (Fig. 3). We evaluated the used HAR method with this dataset in the leave-one-out (LOO) fashion. Table 1 shows the recognition results for each action class. The high-speed of the actions and players' clothes hindered HAR, and similar actions were often mistaken. Its generalization performance is evaluated in [20] against the MSRAction3D dataset with the configuration used in [6]. The used method has an accuracy of 84.1%, surpassing [6] (74.7%), and other nearest neighbors-based methods [23] (63%). However, this accuracy is a bit lower than that of methods with a more costly training, such as support vector machines [24] (88.2%), or convolutional neural networks [25] (94.6%).

We asked 13 participants to evaluate our method. Since the interestingness of the extracted highlights can differ from one user to another, we grouped them into experienced (E) and non-experienced (NE) in Kendo, which would affect the results the most. Group E has 3 users and NE has 10. In order to train the GMM-HMM for highlight extraction, 3 and 5 users from groups E and NE were employed as annotators, and assigned interesting/non-interesting labels to the sub-sequences in the 10 original videos. Each sub-sequence was judged to be interesting if two or more annotators labeled it as interesting. Whereas group E picked sub-sequences with very specific actions (*e.g.*, very fast strikes, decisive strikes, etc.), group NE picked a more general set of actions (*e.g.*, non-decisive strikes, feints, etc.), reaching about twice the number of sub-sequences than group E. Again in the LOO fashion, we trained the GMM-HMM with the labels of 9 videos to generate the summary of the remaining.

Table 2. GMM-HMM performance.

	Annot. E			Annot. NE		
	P	R	F	P	R	F
(A)	0.41	0.44	0.42	0.62	0.76	0.68
(B)	0.39	0.42	0.41	0.62	0.75	0.68
(C)	**0.57**	**0.72**	**0.63**	**0.79**	**0.77**	**0.78**
(D)	0.49	0.64	0.56	0.77	0.75	0.76

4.1. GMM-HMM objective evaluation

We evaluated the performance of our trained GMM-HMM by thresholding $\Pr(e_t = 1|F) > 0.5$, and calculating precision (P), recall (R), and f-score (F) metrics for the extracted sub-sequences. Due to the limitations of the capturing device, in some parts of the original video, one or both players were not recognized. For this reason, we evaluated the performance under these conditions: all sub-sequences (A, B) and only the sub-sequences in which both players' skeleton is tracked (C, D). We also evaluated the difference in performance when the activity measure is used (A, C) or not (B, D). Table 2 shows the results. The best results correspond to the case where both players' skeletons were tracked and activity measure was used (C). The effect of including our activity measure is greater on group E's results. Since group E's annotations included more specific actions, it seems the activity measure helps to discern specific interesting actions among similar HAR results. When comparing groups E and NE, the latter's performance is higher since their annotations contain a broader set of actions.

4.2. Video summary objective evaluation

Our generated summaries are composed of sub-sequences with their estimated labels of interestingness. Human annotators expected that a set of consecutive sub-sequences with interest labels (referred to as a highlights, hereinafter) contain an event in a certain granularity. Therefore, even a single missed sub-sequence in the set may distract viewers. For this, we objectively evaluated our method by modifying the definitions of precision and recall to take into account the completeness of the extracted highlights. We define the completeness criterion for an extracted highlight as the fraction of overlap with its associated highlight from the ground truth annotated by our participants. Associating extracted and ground truth highlights is not trivial, and we did this in a greedy manner, in which the total number of overlapping sub-sequences is maximized. We deemed an extracted highlight as a true positive (TP) if it covers over $C\%$ of the sub-sequences in the associated ground truth highlight. In this experiment, we thresholded $\Pr(e_t|F_T)$ in the range $[0, 1]$ (instead of 0.5 as in section 4.1) to generate summaries of different lengths.

Figure 4 shows the recall-precision curves produced for $C = 50\%, 70\%, 90\%$. Whereas almost all highlights with $C = 70\%$ reached also $C = 90\%$, when reducing C to 50% the number of TP increases significantly. We attribute the presence of incomplete segments to the transition prob-

Table 3. Survey results. Each cell consists of the mean ± standard deviation of the subjective scores.

		Summary type			Length			Video		
		Annot. E	Annot. NE	Clust.	20 s	30 s	40 s	(a)	(b)	(c)
Q1	Grp. E	**3.44±0.67**	3.04±0.72	1.89±0.69	3±0.7	**3.56±0.58**	3.17±0.81	**3.61±0.88**	3.11±0.58	3±0.56
	Grp. NE	**3.63±0.5**	**3.63±0.49**	2.26±0.78	3.58±0.46	**3.75±0.43**	3.57±0.61	**3.9±0.54**	3.75±0.35	3.25±0.33
Q2	Grp. E	**3.33±0.58**	3±0.33	1.37±0.35	2.89±0.62	**3.33±0.21**	3.28±0.49	**3.28±0.57**	**3.28±0.39**	2.94±0.49
	Grp. NE	**3.79±0.53**	3.78±0.3	1.88±0.55	3.53±0.5	**3.92±0.32**	3.9±0.36	**4.1±0.29**	3.8±0.24	3.45±0.45
Q3	Grp. E	**3.33±0.33**	3.11±0.58	1.33±0.29	3.11±0.66	**3.33±0.21**	3.22±0.5	**3.33±0.67**	3.22±0.46	3.11±0.27
	Grp. NE	3.57±0.54	**3.68±0.39**	1.92±0.49	3.38±0.48	**3.77±0.38**	3.72±0.49	**3.88±0.48**	3.65±0.31	3.33±0.45
Q4	Grp. E	4.41±0.57	**4.67±0.33**	2.22±0.58	4.44±0.69	**4.61±0.44**	4.56±0.27	**4.72±0.33**	4.61±0.44	4.28±0.57
	Grp. NE	3.6±0.34	**3.62±0.36**	2.27±0.35	3.47±0.41	**3.8±0.27**	3.57±0.29	**3.88±0.32**	3.52±0.25	3.43±0.3

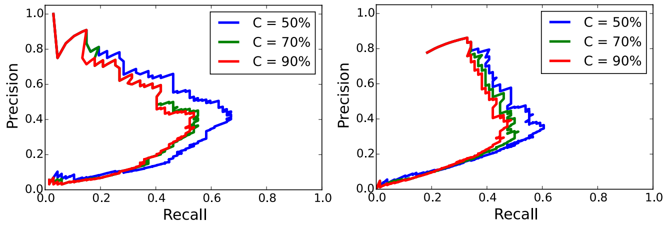


Fig. 4. Recall-precision curves for grp. E (left) and NE (right)

abilities of our GMM-HMM model, which are very low for the *non-interesting to interesting* transition and higher for the *interesting to non-interesting* one. This makes highlights start later and begin earlier than the annotated ground truth. When comparing groups E and NE, the latter's recall shows a higher and more constant number of TPs for different summary lengths, which is consistent with the results shown in section 4.1. We conclude that our method is able to detect very well certain highlights, but others remain incomplete.

4.3. Video summary subjective evaluation

We assessed the quality and usefulness of our video summaries from the users' point of view by means of a survey. All 13 participants watched the video summaries that, for $C = 70\%$, gave the (a) maximum, (b) median, and (c) minimum f-scores averaged for groups E and NE in the previous section, as well as their corresponding original video. We also used different summary lengths $L = 20$, 30, and 40 s, to see how the length affects viewers' perception. For comparison, besides the summaries created with groups E and NE annotations, we also evaluated video summaries based on the k-means clustering algorithm as a baseline, in which clustering was performed on our HAR features. As a result, every participant watched 27 summaries.

We asked participants (Q1) if each summary showed an entire action from beginning to end, (Q2) if each summary was interesting, (Q3) if the participant got an insight on the original video by watching the summary, and (Q4) if the summary was not redundant. Table 3 shows the results for each question. Answers are averaged for group E and NE separately and grouped by the summary type, length, and video. The latter two cover the answers for summaries created with annotations E and NE together. By looking at the first row, the answers to Q1 show that users were satisfied with the completeness of our summary. Q2 and Q3 also show the user's satisfaction, although group E's rating is slightly lower than group NE's. This is probably because the experienced participants wanted to see all interesting highlights in the summary, but some were missing. The inexperienced participants did not have such a firm predilection. In Q4, group NE found the summaries more redundant than group E, in a way that group NE preferred watching also non-active segments before the action starts for a better understanding of the context.

When comparing summary types, it can be observed that the clustering-based baseline has the lowest scores for all the questions. Overall, group E rated the summaries created with their annotations higher, except in Q4. For group NE, the difference between summaries generated with their annotations or with group E's is not noticeable. Regarding length, 30 second summaries obtained the best evaluation for all questions and user groups. We consider the reason is that 20 second summaries contained some incomplete highlights that were filled in the 30 second ones, but in the 40 second summary, newly added highlights were incomplete. The summary for video (a) was ranked higher for all questions and both groups, which is coherent since it has the highest f-score.

Some participants in group NE commented the usefulness of our method to extract highlights based on actions, and the time they can save by watching the summary instead of the whole video. Group E stated that in Kendo it is important to observe the actions after hitting the opponent as well (even if they are not interesting) in order to decide if it was a good hit. However, when creating a summary for a given length, our method gives priority to extracting new interesting highlights rather than adding less interesting sub-sequences to the existing ones. All our participants preferred watching longer highlights rather than a larger number of them.

5. CONCLUSIONS

In this paper we have presented a novel method for generating video summaries with highlights of personal sports video by using HAR, which is used to train a highlights model based on viewers' opinion on which sections of the original video were interesting. Our experiments and the positive responses from the survey showed that our method was able to successfully extract highlights using HAR, despite our HAR was not perfect. We believe the reason is that our method does not directly rely on HAR results, but on its intermediate outputs, which can leverage the ambiguity among different action classes. Although we experimented with only one type of sport, *i.e.*, Kendo, our method is applicable to other similar sports. As future work, we will investigate a way to include the context into highlights. In order to support our results, we need more participants in survey as well as annotators. More sophisticated models for highlight extraction, *e.g.*, recurrent neural networks, would be another research direction.

Acknowledgements This work is partly supported by JSPS KAKENHI No. 16K16086.

6. REFERENCES

[1] B. T. Truong and S. Venkatesh, "Video abstraction: A systematic review and classification," *ACM Trans. Multimedia Computing, Communications, and Applications*, vol. 3, no. 1, pp. 3:1–3:37, 2007.

[2] J. Choi, W. J. Jeon, and S. C. Lee, "Spatio-temporal pyramid matching for sports videos," in *Proc. ACM Int. Conf. Multimedia Information Retrieval*, 2008, pp. 291–297.

[3] B. Li and M. I. Sezan, "Event detection and summarization in sports video," in *Proc. IEEE Workshop on Content-based Access of Image and Video Libraries*, 2001, pp. 132–138.

[4] N. Nitta, Y. Takahashi, and N. Babaguchi, "Automatic personalized video abstraction for sports videos using metadata," *Multimedia Tools and Applications*, vol. 41, no. 1, pp. 1–25, 2009.

[5] P. V. K. Borges, N. Conci, and A. Cavallaro, "Video-based human behavior understanding: a survey," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 23, no. 11, pp. 1993–2008, 2013.

[6] W. Li, Z. Zhang, and Z. Liu, "Action recognition based on a bag of 3d points," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition Workshops*, 2010, pp. 9–14.

[7] A. Ekin, A. M. Tekalp, and R. Mehrotra, "Automatic soccer video analysis and summarization," *IEEE Trans. Image Processing*, vol. 12, no. 7, pp. 796–807, 2003.

[8] H. Pan, P. Van-Beek, and M. I. Sezan, "Detection of slow-motion replay segments in sports video for highlights generation," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, 2001, vol. 3, pp. 1649–1652.

[9] M. Chen, S. C. Chen, M. L. Shyu, and K. Wickramaratna, "Semantic event detection via multimodal data mining," *IEEE Signal Processing Magazine*, vol. 23, no. 2, pp. 38–46, 2006.

[10] W. H. Cheng, Y. Y. Chuang, Y. T. Lin, C. C. Hsieh, S. Y. Fang, B. Y. Chen, and J. L. Wu, "Semantic analysis for automatic event recognition and segmentation of wedding ceremony videos," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 18, no. 11, pp. 1639–1650, 2008.

[11] X. S. Hua, L. Lu, and H. J. Zhang, "Optimization-based automated home video editing system," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 14, no. 5, pp. 572–583, 2004.

[12] R. W. Lienhart, "Dynamic video summarization of home video," in *Proc. SPIE Electronic Imaging*, 1999, pp. 378–389.

[13] D. Gatica-Perez, A. Loui, and M. T. Sun, "Finding structure in home videos by probabilistic hierarchical clustering," *IEEE Trans .Circuits and Systems for Video Technology*, vol. 13, no. 6, pp. 539–548, 2003.

[14] R. Laganière, R. Bacco, A. Hocevar, P. Lambert, G. Païs, and B. E. Ionescu, "Video summarization from spatio-temporal features," in *Proc. ACM TRECVid Video Summarization Workshop*, 2008, pp. 144–148.

[15] C. Y. Chen, J. C. Wang, J. F. Wang, and Y. H. Hu, "Motion entropy feature and its applications to event-based segmentation of sports video," *EURASIP Journal on Advances in Signal Processing*, vol. 2008, pp. 1–8, 2008.

[16] W. T. Peng, W. T. Chu, C. H. Chang, C. N. Chou, W. J. Huang, W. Y. Chang, and Y. P. Hung, "Editing by viewing: automatic home video summarization by viewing behavior analysis," *IEEE Trans. Multimedia*, vol. 13, no. 3, pp. 539–550, 2011.

[17] G. Zhu, C. Xu, Q. Huang, W. Gao, and L. Xing, "Player action recognition in broadcast tennis video with applications to semantic analysis of sports game," in *Proc. ACM Int. Conf. Multimedia*, 2006, pp. 431–440.

[18] G. Ballin, M. Munaro, and E. Menegatti, "Human action recognition from RGB-D frames based on real-time 3D optical flow estimation," in *Proc. Biologically Inspired Cognitive Architectures*, pp. 65–74. 2013.

[19] Z. Zhang, "Microsoft Kinect sensor and its effect," *IEEE MultiMedia*, vol. 19, no. 2, pp. 4–10, 2012.

[20] A. Tejero-de-Pablos, Y. Nakashima, N. Yokoya, F. J. Díaz-Pernas, and M. Martínez-Zarzuela, "Flexible human action recognition in depth video sequences using masked joint trajectories," *EURASIP Journal on Image and Video Processing*, to be published.

[21] L. Rabiner and B. H. Juang, "Fundamentals of speech recognition," 1993.

[22] T. K. Moon, "The expectation-maximization algorithm," *Signal processing magazine*, vol. 13, no. 6, pp. 47–60, 1996.

[23] B. B. Amor, J. Su, and A. Srivastava, "Action recognition using rate-invariant analysis of skeletal shape trajectories," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 38, no. 1, pp. 1–13, 2016.

[24] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2012, pp. 1290–1297.

[25] P. Wang, W. Li, Z. Gao, J. Zhang, C. Tang, and P. Ogunbona, "Deep convolutional neural networks for action recognition using depth map sequences," *arXiv preprint arXiv:1501.04686*, 2015.