Salient Time Slice Pruning and Boosting for Person-Scene Instance Search in TV Series

Zheng WANG[1,a)] Fan YANG[2,b)] Shin'ichi SATOH[1,c)]

Abstract

It is common that TV audiences want to quickly browse scenes with certain actors in TV series. Since 2016, the TREC Video Retrieval Evaluation (TRECVID) Instance Search (INS) task has started to focus on identifying a target person in a target scene *simultaneously*. We name this kind of task as P-S INS (Person-Scene Instance Search). We find that person and scene INS modules may suppress each other in some situations. Luckily, video shots are arranged in chronological order. We extend our focus from *time point* to *time slice*. Through detecting salient time slices, we prune the dataset. Through evaluating the importances of salient time slices, we boost the aggregation results. Extensive experiments on the large-scale TRECVID INS dataset demonstrate the effectiveness of the proposed method.

1. Introduction

TV audiences will appreciate a system that can show him someone of interest in certain scenes, for example, "Sheldon stays in Amy's house" in the TV series "Big Bang Theory". TREC Video Retrieval Evaluation (TRECVID) takes notice of such system. Since 2016, the INstance Search (INS) task of TRECVID has started to ask participants to search out shots, which contain a certain person appearing in a certain place [1]. We name this kind of instance search task, aiming at identifying a target person in a target scene, as Person-Scene INstance Search (P-S INS). Although the general INS task, focusing on a single target *independently*, has already been well studied [3], the P-S INS task, aiming at doing the retrieval job based on two different kinds of instances *simultaneously*, is challenging and just catching up (Fig. 1).

In TV series, person may appear in any angle or corner of the scene. Person's appearance always varies, and scene's viewpoint always changes. It makes the content of P-S instance pair is ever-changing. Hence, a P-S instance pair must not be taken as an integral whole. Due to wildly different characteristics between person instances and scene instances, person INS and scene INS always exploit different technology roadmaps. Most of existing methods utilize person INS and scene INS modules apart, search for target persons and scenes respectively, and finally combine the results together to generate ranking lists [1], [2]. However, **person INS module and scene INS module are not always effective at the same time**, or they often suppress each other in some situations. Fig. 2 gives some examples. Person INS module may be not effective because faces are not detected when they are non-front or occluded (Fig. 2(a) and Fig. 2(b)). Scene INS module may be also not effective because of blur and low light (Fig. 2(c) and Fig. 2(d)). In a wide-angle view, scene INS module performs well, but person INS module is constrained because persons (in particular faces) in this kind of scenes are very small (Fig. 2(e) and Fig. 2(f)). Scenes may be blocked by persons, which makes scene INS module suppressed (Fig. 2(g) and Fig. 2(h)). Consequently, directly aggregating the results shot by shot will not obtain satisfactory results.

Fig. 1 Examples for general INS task (top) and P-S INS task (bottom). The examples are selected from the TV series *Eastenders*. Programme material copyrighted by BBC.

Luckily, video shots are arranged in chronological order. If one target person/scene is captured in a single shot, it is very likely that this person/scene will also appear in the neighbor shots in the time-line. So even though person INS and scene INS modules perform not good enough due to the hard conditions described above, we may get some high person/scene scores within a couple of consecutive shots, because the target person will stay in the target scene for a little while. It should be mentioned that the similarity between the query topic and each shot is used to generate ranking list, and we use person/scene score to stand for the similarity calculated by person/scene INS module. Inspired by the video consecutiveness, we extend the focus from *time point* (single video shot) to *time slice* (multiple consecutive

[1] National Institute of Informatics
[2] The University of Tokyo
[a)] wangz@nii.ac.jp
[b)] yang@nii.ac.jp
[c)] satoh@nii.ac.jp

Fig. 2 Some hard shots.

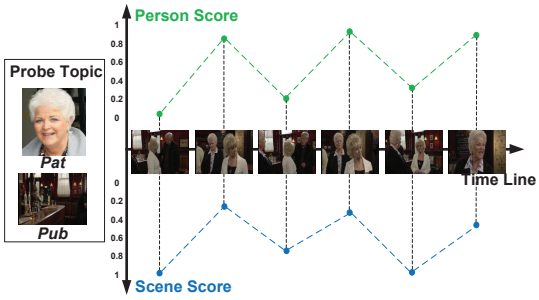


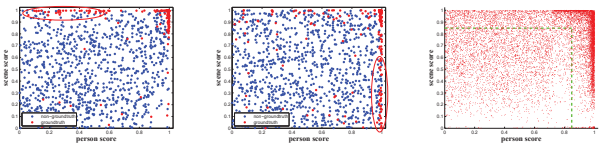
Fig. 3 An example for consecutive shots in a time slice.

video shots). Fig. 3 shows six consecutive shots as an example. Among these shots, person and scene scores are not high at the same time. If the scores are combined shot by shot directly, we will obtain unsatisfied performance. Nevertheless, if the maximum score of these consecutive shots is high, it tells us that the person/scene must will appear in these shots with high probability. The time slice, composed by the shot and its neighbor shots in the time-line, can be taken as an indicator of the possibility of person's/scene's appearance. We name this kind of slice in time-line as Salient Time Slice (STS).

In this paper, we study to detect STS. The STS is employed to not only prune P-S INS data, but also boost P-S INS results. The contribution of this paper is as follows: 1) We observe that person INS module and scene INS module are in a dilemma for P-S INS task. Groundtruth shots mainly lie in a rotated "L" shape area of P-S score coordinate plane, rather than the right top corner of the plane. 2) Considering the consecutiveness, we extend the focus from time *point* to time *slice*, and design the STS pruning and boosting method for INS results aggregation. 3) Extensive evaluations on the large-scale dataset show the superiority of our method, even though it is very simple.

2. Investigation on P-S INS Scores

We conduct preliminary experiments to investigate the normalized P-S score distribution. We use the TRECVID INS dataset to do the experiments. We first exploit basic person and scene INS modules to obtain the initial scores respectively. The detail information about the dataset and the INS modules are described in the experiment section. Fig. 4(a) and Fig. 4(b) demonstrate the shot score distributions of two probe topics, respectively in 2016 and 2017. We randomly select 200 groundtruth shots and 1000 non-groundtruth shots for each probe topic. As the circled area of Fig. 4(a) indicates, a lot of groundtruth shots hold a high scene INS score, while a relative low person INS score. The combination of these person and scene scores will not be



(a) 9170 (b) 9210 (c) Groundtruth

Fig. 4 Shot score distributions. The red points stand for the groundtruth shots, and the blue points stand for the non-groundtruth shots.

high enough. Also, as the circled area of Fig. 4(b) indicates, there are also a lot of groundtruth shots with high person INS score and relative low scene INS score. **It indicates that the dilemma for P-S INS task exists**. In addition, we draw all the groundtruth shots in the P-S score coordinate plane, as Fig. 4(c) shows. We can find that the groundtruth shots mainly lie in a rotated "L" shape area of the plane. If a shot lies in this area, it is more likely to be a groundtruth shot. We coordinate a rotated "L" shape area, whose scene score and person score are both over 0.84.

We set a new person/scene score for each shot by max-pooling the scores of its neighborhood shots in the time-line. In this way, we push those shots to the rotated "L" shape area. After maxpooling, we re-evaluate the distribution in this area. The area contains 92.04% and 96.47% groundtruth shots respectively in 2016 and in 2017. It is reasonable to pay our retrieval attention only to the shots with high scene score or person score, because these shots stand a good chance of being groundtruth shots and can help us to find more target shots.

3. STS Pruning and Boosting

The overall scheme of our approach is shown in Fig. 5. Given a probe P-S instance topic, the person and scene INS modules generate two score lists. Each shot contains several keyframes. Each keyframe of the shot will get its initial score from the module. We choose the maximum one to represent the initial shot score. After calculating the scores of all 471,523 shots, we normalize the person and scene INS scores for each P-S instance probe topic. For each probe topic, each test shot i gets two normalized scores (p_i, s_i) respectively based on the initial scores of person and scene INS modules, where $p_i, s_i \in [0, 1]$, p_i is for the person INS, and s_i is for the scene INS. If the shot gets a high similarity by person (scene) INS module, p_i (s_i) will be a high score. The following parts of the framework are divided into two steps: STS pruning and STS boosting.

STS Pruning

For the person INS branch, we utilize neighborhood max-pooling to generate a slice score sp_i for each shot. The score is denoted as: $sp_i = Max(p_{i-K}, ..., p_i, ..., p_{i+K})$, where K stands for the number of neighbors in time-line. In this paper, we set $K = 8$, because it is found that in TV series, an activity or a dialogue often maintain in a scene for around 15 video shots. Generally, we consider that if the slice score is larger than a high threshold score ρ_p, the time slice will be the salient time slice. The threshold score ρ_p is decided

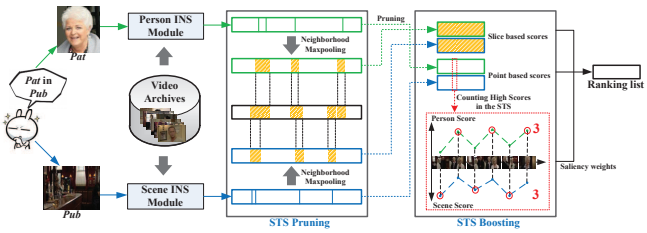


Fig. 5 The framework of our approach.

by the top scores for different topics. To this end, we remain the shots, whose slice scores are larger than ρ_p. For the test shots in STS, we collect shot IDs to form a subset as: $SP = \{i|i \in [1, N], sp_i > \rho_p\}$, where $N = 471,523$ is the total number of gallery shots.

For the scene INS branch, we follow the same manner. The slice score ss_i for each shot is denoted as: $ss_i = Max(s_{i-K}, ..., s_i, ..., s_{i+K})$. We can also collect shot IDs to form a subset based on slice score ss_i. $SS = \{i|i \in [1, N], ss_i > \rho_s\}$, where ρ_s stands for the threshold scene score for selecting salient slices. Finally, the total STS is the union of these two subsets.

$$STS = SP \bigcup SS = \{i|i \in [1, N], sp_i > \rho_p || ss_i > \rho_s\}. \quad (1)$$

As Fig. 5 shows, after STS detection, the proposed framework prunes the data, by remaining the shots in STSs. Actually, the selected shots in STSs lie in a kind of rotated "L" shape area of the plane described above.

STS Boosting If a time slice holds more high person scores and scene scores, the corresponding shot will be more likely to be the groundtruth. We exploit this idea to combine and boost the results. First of all, we evaluate each shot in STS by the number of high scores around its neighbors. To this end, for each shot $i \in STS$, we count the number of high scores respectively in person INS branch and scene INS branch. Suppose the quantities are respectively N_i^p and N_i^s for the shot i. We count the number of high scores using the following two equations. The function $Times()$ counts the number of high scores. Among the shot and its neighbors, if a score is higher than the threshold, the count will add one. Then, we define the saliency weight for the corresponding shot as $w_i = N_i^p + N_i^s$.

$$\begin{aligned} &N_i^p = Times(p_j > \rho_p) \\ &\quad \text{s.t.} \quad j \in [i-K, i+K] \quad \&\& \quad j \in STS \end{aligned} \quad (2)$$

$$\begin{aligned} &N_i^s = Times(s_j > \rho_s) \\ &\quad \text{s.t.} \quad j \in [i-K, i+K] \quad \&\& \quad j \in STS \end{aligned} \quad (3)$$

With STSs and their saliency weights, we calculate the final scores. Considering that the initial score accounts for the effectiveness of person/scene INS module, and the slice score compensates for the dilemma, we combine them together. For each shot, we calculate its final score as:

$$\begin{aligned} &score_i = [\alpha * (p_i + s_i) + (1-\alpha) * (sp_i + ss_i)] \times e^{\beta * w_i} \\ &\quad \text{s.t.} \quad i \in STS, \quad 0 \leqq \alpha \leqq 1, \quad \beta > 0 \end{aligned} \quad (4)$$

Here, α denotes the factor to balance the contribution of the slice and the shot point itself, and β denotes the impact of STS boosting. In this paper, we set $\alpha = 0.05, \beta = 0.001$. Finally, when we obtain all the scores in STSs, we rank them to generate the final ranking list.

4. Experiments

4.1 Dataset and Basic INS Modules

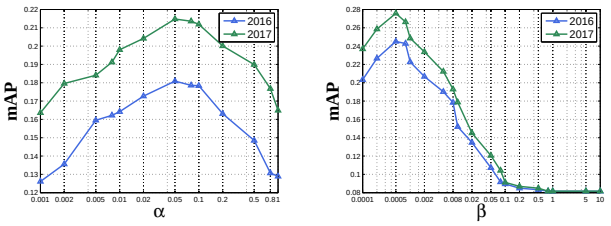
Dataset. We select the TRECVID INS dataset as the experimental dataset [2]. The dataset uses 464 hours of the BBC soap opera *EastEnders*, which is divided by the BBC into 471,523 shots, about 5 million images to be used as the unit of retrieval. In 2016 and 2017, INS task respectively chose representative sample of 30 probe topics and ask participants to search video shots [1], [2]. Each probe topic is made up of a selected pair of person and scene. For each probe topic, we should search and return up to the 1000 shots most likely containing the person and scene of the topic. Following the standard evaluation criteria, we test the performance by the metric mean Average Precision (mAP) [8].

Basic person INS module. We utilized the facial appearance as the retrieval cue. The module mainly consists of two steps, one is face detection, and the other is face identification. For face detection, we adopted the Scale-Adaptive Deconvolutional Regression (SADR) network [10]. For person INS module, we built our own face reference set. We collected face images as more as possible through *Bing*. Finally, the reference set includes 815 face images. In particular, each target person holds multiple face images, and each non-target only contains single face image. The score of each shot is obtained by maxpooling the scores of images within the shot.

Basic scene INS module. Scene INS in present is based on global or local views. On one hand, based on a deep learning system (ResNets [5]), we took the output of a pre-trained CNN as the global scene feature. We adopted the Facebook's 152-layer model [4], and the output of the model was denoted as the global scene feature. On the other hand, based on a hand-crafted system, such as BoW, through identifying typical objects in certain scenes, we also seek out target scenes indirectly. Following [6], several different landmark objects were selected for each topic scene. From global and local views of scene, we got two scene INS results for each topic scene, and fused them for the scene INS result.

4.2 Investigation on the impact of α and β

First, we set β as zero, and observe the variation of mAP value as α changes. In this situation, no boosting weight is introduced. Fig. 6(a) shows the results of both TRECVID INS dataset in 2016 and 2017. From the figure, we can find that the mAP value goes up and down, as α changes from 0.001 to 1. And the changing trend is the same for both two



(a) The impact of α (b) The impact of β

Fig. 6 Investigation on the impact of α and β.

years' topics. So the effectiveness of point score and slice score are different, and the parameter α can balance the contributions of these two kinds of scores. However, we can also find that when $\alpha = 0.05$, the mAP value reaches at maximum. Hence, more weight should be assigned to the slice score, and slice score is important than point score. Second, we set $\alpha = 0.05$, and observe the variation of mAP value as β changes. Fig. 6(b) shows the results of both TRECVID INS dataset in 2016 and 2017. From the figure, we can find that the mAP value goes up and down, as β changes from 0.0001 to 10. And the changing trend is the same for both two years' topics. We can also find that when $\beta > 1$, the mAP value will not change. In that situation, the strength of boosting is too large, original shot scores will have no contribution to the final result. However, when we choose a suitable β, such as $\beta \in [0.0001, 0.002]$, the boosting saliency weight will be very useful and effective.

4.3 Evaluation on the Proposed Method

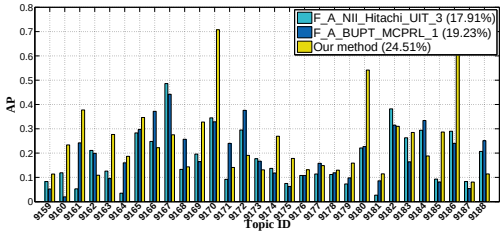
From the Table. 1, we find that the combination of point based and slice based scores can get a high performance. Meanwhile the saliency weight makes sense, which helps our final results make a considerable improvement. That is to say, changing focus from time point to slice is necessary, and STS saliency evaluation is an effective way for boosting P-S INS task.

Table 1 Evaluation in term of mAP on the proposed method with different strategies.

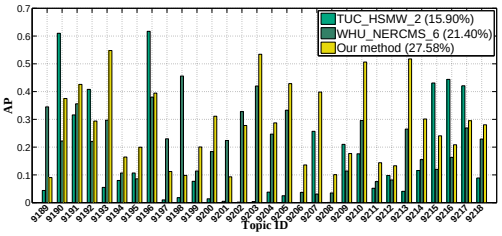
Different strategies	2016	2017
$p_i + s_i$	12.89%	16.49%
$\alpha * (p_i + s_i) + (1 - \alpha) * (sp_i + ss_i)$	18.03%	21.34%
$score_i$	24.51%	27.58%

4.4 Comparison with other Methods

To prove the effectiveness of our method, we choose representative methods, which are excellent but have not many tricks. We chose two representative methods to make the comparison for each year. In Fig. 7, we drew the AP values topic after topic and listed the mAP results at the right corner. In 2016 topics, we selected the F_A_NII_Hitachi_UIT_3 [7] and F_A_BUPT_MCPRL_1 [9] methods. From the Fig. 7(a), we can find that our method outperforms the others in 18/30 topics. Some topics does not performs better than the other methods, because our scene INS module performs not well enough. For example, the scene of topic 9167, 9172, 9177 and 9188 is *Living Room1*. In 2017 topics, we selected the WHU_NERCMS_6



(a) AP and mAP in TRECVID INS 2016



(b) AP and mAP in TRECVID INS 2017

Fig. 7 The comparisons with the other methods by AP for each topic and mAP.

and TUC_HSMW_2 [1] methods. From the Fig. 7(b), we can find that 17 topics with our method achieve higher performance. Due to imperfect effectiveness of our scene INS module, some topics do not perform better than the other methods. For example, the scene of topic 9192, 9196 and 9215 is *Cafe1*. Although we exploit different person and scene INS modules, compared with the other methods, our method still gets considerable performance.

5. Conclusion

We find that there is a negative correlation for the person and scene INS results in P-S INS task. Inspired by the consecutiveness of TV series, we extend the focus from time point to time slice. We first prune the dataset by detecting STS in time-line, and then aggregate and boost the results in STS. The method is very simple and effective.

References

[1] Awad, G., Butt, A. and Fiscus, J.: Trecvid 2017: Evaluating ad-hoc and instance video search, events detection, video captioning and hyperlinking, *Proceedings of TRECVID* (2017).

[2] Awad, G., Fiscus, J. and Michel, M.: Trecvid 2016: Evaluating video search, video event detection, localization, and hyperlinking, *Proceedings of TRECVID* (2016).

[3] Awad, G., Kraaij, W., Over, P. and Satoh, S.: Instance search retrospective with focus on TRECVID, *International journal of multimedia information retrieval* (2017).

[4] Gross, S. and Wilber, M.: Training and investigating residual nets, *Facebook AI Research* (2016).

[5] He, K., Zhang, X., Ren, S. and Sun, J.: Deep residual learning for image recognition, *CVPR* (2016).

[6] Lan, J., Chen, J., Wang, Z., Liang, C. and Satoh, S.: PS Instance Retrieval via Early Elimination and Late Expansion, *ACM MM Workshop* (2017).

[7] Le, D.-D., Phan, S. and Satoh, S.: NII-HITACHI-UIT at TRECVID 2016, *TRECVID Workshop* (2016).

[8] Salton, G. and Harman, D.: *Information retrieval*, John Wiley and Sons Ltd. (2003).

[9] Zhao, Z., Wang, M. and Xiang, R.: Bupt-mcprl at trecvid 2016, *TRECVID Workshop* (2016).

[10] Zhu, Y., Wang, J., Zhao, C., Guo, H. and Lu, H.: Scale-adaptive deconvolutional regression network for pedestrian detection, *ACCV* (2016).